ABSTRACT

Title of Dissertation: EFFECTS OF ACOUSTIC PERCEPTION OF GENDER ON NONSAMPLING ERRORS IN TELEPHONE SURVEYS

Susan Kenney McCulloch, Doctor of Philosophy, 2012

Dissertation Directed by: Dr. Frauke Kreuter, Joint Program in Survey Methodology

Many telephone surveys require interviewers to observe and record respondents' gender based solely on respondents' voice. Researchers may rely on these observations to: (1) screen for study eligibility; (2) determine skip patterns; (3) foster interviewer tailoring strategies; (4) contribute to nonresponse assessment and adjustments; (5) inform post-stratification weighting; and (6) design experiments. Gender is also an important covariate to understand attitudes and behavior in many disciplines. Yet, despite this fundamental role in research, survey documentation suggests there is significant variation in how gender is measured and collected across organizations. Variations of collecting respondent gender may include: (1) asking the respondent; (2) interviewer observation only; (3) a combination of observation aided by asking when needed; or (4) another method. But what is the efficacy of these approaches? Are there predictors of observational errors? What are the consequences of interviewer misclassification of respondent gender to survey outcomes? Measurement error in interviewer's observations of respondent gender has never been examined by survey methodologists.

This dissertation explores the accuracy and utility of interviewer judgments specifically with regard to gender observations. Using the recent paradata work and linguistics literature as a foundation to explore acoustic gender determination, the goal of

my dissertation is to identify implications for survey research of using interviewers' observations collected in a telephone interviewing setting.

Organized into three journal-style papers, through a survey of survey organizations, the first paper finds that more than two-thirds of firms collect respondent gender by some form of interviewer observation. Placement of the observation, rationale for chosen collection methods, and uses of these paradata are documented. In paper two, utilizing existing recording of survey interviews, the experimental research finds that the accuracy of interviewer observations improves with increased exposure. The noisy environment of a centralized phone room does not appear to threaten the quality of gender observations. Interviewer and respondent level covariates of misclassification are also discussed. Analyzing secondary data, the third paper finds there are some consequences of incorrect interviewer observations of respondents' gender on survey estimates. Findings from this dissertation will contribute to the paradata literature and provide survey practitioners guidance in the use and collection of interviewer observations, specifically gender, to reduce sources of nonsampling error.

Effects of Acoustic Perception of Gender on Nonsampling Errors in Telephone Surveys

by

Susan Kenney McCulloch

Dissertation submitted to the Faculty of the Graduate School of the University of Maryland, College Park in partial fulfillment of the requirements for the degree of Doctor of Philosophy

2012

Advisory Committee:

Dr. Frauke Kreuter, Chair/Adviser
Dr. Frederick Conrad
Dr. Lee Miringoff
Dr. Stanley Presser
Dr. William Idsardi, Dean's Representative

Dedication

This dissertation is dedicated to my parents, Nelson and Laurie Kenney. It is a product of their unconditional love, constant support, and tireless motivation. Mom and Dad, thank you for making every one of my dreams come true, for opening up worlds of opportunity, and for making, what I thought to be, the impossible happen. You inspire me to achieve. Once again, you have cheered (and dragged) me across the finish line. I can only hope this work is an improvement from *Green Beans*.

Acknowledgments

Many people endured this journey along with me. I am forever appreciative for their support, encouragement, and guidance.

I am indebted to my adviser, Frauke Kreuter. In a single car ride from Vancouver to Whistler Mountain, she changed my life—fostering this work and committing to the accomplishment of my Ph.D. Frauke, you are a brilliant mentor, colleague, and friend. I am grateful for your persistence and devotion to my success. Thank you for challenging me, drawing upon my strengths while demanding I improve my weaknesses, believing in my ability, and for the opportunities you've presented me.

I am extremely grateful to my committee members, Stanley Presser, Fred Conrad, Bill Idsardi, and Lee Miringoff who provided continual feedback and constant support. Thank you for helping me bridge the practical relevance of this work with academic rigor. I am honored to have worked with such a prestigious team of scholars.

To the JPSM faculty and staff, especially Katharine Abraham and Rick Valliant, thank you for your confidence and patience in the completion of my degree. The technical and theoretical training I received in this program is invaluable. I am grateful for the network of accomplished colleagues JPSM has provided.

I am eternally indebted to my Marist College Poll mentors Lee Miringoff and Barbara Carvalho, without whom this dissertation would not be possible. They provided not only all of the necessary data for this research (both primary data collection and datasets for secondary analysis), but years of encouragement and countless hours of feedback, resources...and employment! Lee and Barb, thank you for bringing me into the world of survey research and public opinion, for your confidence, determination,

flexibility, compassion, and friendship. To my colleagues Mary Griffith, Stephanie Calvano, Jaime Lynne Bishop, Alicia Colomer, Daniela Charter, and Natalie Jackson: thank you, all, for your patience, support, and for enduring the "D" along with me. Does this help get me into *The Club*? Also, to Mike Conte: I'm so happy you can now include *Praat* in your list of academic achievements. Thank you for your hard work.

This work was strengthened by the feedback and thorough review of many other JPSM students, colleagues, and friends, especially: Stephanie Eckman, Karen Bryce, Jenna Fulton, Frank Ferrari, Samrat Bose, Jose Benki, and Claire Delaney. I am grateful for each of their comments, questions, technical help, and brainstorming of ideas. Thank you!

This dissertation has been a labor of love for my family, especially, Laurie and Nelson Kenney, Jason, Julia, and Carmella Kenney, Uncle Frank and Aunt Janet Ferrari, and Uncle Ed Delaney. Thank you, all, for having more faith in me than I had in myself, for making me laugh when I lost my sense of humor, and for providing Oreos and home-cooked meals. My friends, especially Susan Appe and Laura Brauer, provided endless support and escapes when I needed a break. Greg, thank you for helping make this possible and for your encouragement. And, Grandma, you were a great listener when I simply needed to talk about how overwhelmed I felt. I am so lucky to have been surrounded by such love throughout these (many!) years. Payback starts now.

Table of Contents

List of Tables

List of Figures

1. Introduction

Gender is an important covariate to understand attitudes and behavior in many disciplines. It enlightens our discussions of topics such as, politics, psychology, health, economics, finance, social norms, and social activities. It is a frequent experimental variable in survey work. The similarities and/or differences between men and women are usually considered in research results. In addition, respondent gender often informs the logistics of studies. For instance, depending upon the goals of the research, respondent gender may be used by survey organizations to: (1) screen for study eligibility; (2) determine skip patterns; (3) foster interviewer tailoring strategies; (4) contribute to nonresponse assessment and adjustments; and (5) inform post-stratification weighting.

Despite this fundamental role in research, survey methodologists have overlooked how to best collect this respondent data while other demographics, such as race and income have received a lot of attention in the question construction literature (e.g., Davern et al., (2005)). Survey documentation (examples are later below) suggests there is significant variation in how gender is measured and collected across organizations and from study to study. Variations of collecting respondent gender may include: (1) asking the respondent (e.g., *Are you male or female?*); (2) interviewer observation only (e.g., *Interviewer: Record the gender of the respondent*.); or, (3) some combination of both approaches (e.g., *Interviewer: Record the gender of the respondent. Ask only if not obvious*.).

What is the efficacy of various approaches to collecting respondent gender? Are there situational predictors, such as the length of exposure to the voice, and interviewer or respondent level covariates of classification error? To what extent is the potential

variation in the quality of gender classification problematic for survey practitioners? Are there disparities in final survey estimates when using interviewer observations versus by respondent reports to analyze data?

Increasingly, researchers have charged interviewers with not only making observations, but documenting their judgments in the form of paradata (Couper, 1998). These interviewer judgments have been used for responsive design decisions (Groves and Heeringa, 2006) and nonresponse adjustment (Kreuter & Casas-Cordero, 2010; West, 2010a; Casas-Cordero, 2010). However, the quality of these judgments and observations, and thus their utility, has only recently started to be investigated. Existing work focuses on face-to-face surveys. Thus, the quality of observational data in telephone surveys is largely unknown.

Telephone surveys have a narrower range of cues to assist interviewers in their judgments than in-person methods. The setting in which telephone interviewers are asked to make gender judgments—together with limited acoustic information to disentangle distinctions in acoustic cues—is likely to increase survey errors. When used for screening for survey eligibility or even filtering for survey questionnaire logic, gender assessments are likely to be made quite early in the survey, sometimes after only hearing a few words. While it seems straightforward that interviewers can use a variety of vocal cues to discriminate between men and women's voices, linguists have demonstrated that listeners may fail to make correct gender classification. This suggests at least some level of error in interviewer's observations of respondent gender.

When interviewer observations are used to determine survey screening and logic, the result may include random misclassification of gender which could, disrupt the

intended questionnaire flow, and decrease the efficiency of nonresponse adjustments. Moreover, when such misclassification is systematic, bias may lead to errors in final survey estimates.

As part of their discussion of race and gender on interviewer effects, Callegaro et al. (2005) state—without referring to a specific study—"Interviewers guess the gender of the respondents all the time since they are trained not to ask about the gender of the respondent unless they are uncertain" (pg. 3816). Possible arguments for using interviewer observations rather than asking a respondent to report their gender include: (1) a belief that such questions may be insensitive or offensive; (2) a perception that omission saves time or reduces survey length; or (3) an assumption that the question is unnecessary given that the answer is generally obvious.

The Behavioral Risk Factor Surveillance System (BRFSS) is an example of a federally-sponsored national telephone survey which, at times, collects respondent gender solely by observation. According to the 2011 questionnaire documentation,[1] for households with more than one person, the gender of each member is enumerated. However, for single person households, interviewers are instructed to record gender by observation and ask only "if necessary." Given this is part of the household selection process, gender assessments are made very early in the questionnaire. This data point informs survey logic, weighting, and analysis.

As another example, the Health Information National Trends Survey (HINTS) used a combination of interviewer observation and respondent report to collect gender

[1] http://www.cdc.gov/BRFSS/questionnaires/pdf-ques/2011brfss.pdf

data. This national survey, commissioned by The National Cancer Institute from 2003-2007, provided data about access to cancer information, perceptions of cancer risks, and patterns of health care needs. As per the 2005 survey documentation[2], early in the telephone survey, interviewers observe the respondents' gender. This judgement, however, was aided by a first name. Only when it was "not obvious" were interviewers instructed to ask whether they were speaking with a man or woman. The interviewer observation then determined survey skip patterns, as several batteries of questions were only asked of females (e.g., breast and cervical cancer modules) while others targeted men (e.g., prostate cancer module). In addition, HINTS' interviewer gender observations were used for post-stratification weighting. HINTS' approach presumes a correlation between certainty and accuracy of listener observations. Yet, the validity of this assumption has not been established. Some research provides support (Bull and Clifford 1984), but other studies do not (Goggin et al., 1991; Hollien et al., 1982; Yarmey, 1995).

While the BRFSS and HINTS examples highlight a hybrid method of collecting respondent gender (interviewer observation combined with respondent report based on the interviewer's perception of the certainty of his/her judgment), many organizations collect respondent gender exclusively by interviewer judgment. For example, at the end of Cornell University's National Social Survey (CNSS), the questionnaire instructs:[3]

> *Interviewer: Record the respondent's gender but don't read this statement or the options:*
> *Male*
> *Female*
> *Do not know*

---

[2] http://hints.cancer.gov/docs/HINTS_2005_Instrument-English.pdf

[3] http://www.sri.cornell.edu/sri/files/cnss/2010/reports/CNSS2010questionnaire.pdf

Refused

Organized in three journal-style papers, this dissertation explores the accuracy and utility of interviewer judgments specifically with regard to gender observations. Using the recent paradata work and linguistics literature as a foundation to explore acoustic gender determination, the goal of my dissertation is to identify implications of using interviewers' observations collected in a centralized telephone interviewing setting for survey research. To do so, I collect both observational and experimental data and analyze both primary and secondary data sources.

The first paper documents how respondent gender is determined in surveys, including where (placement) this data is collected in the questionnaire, and the ways in which firms use interviewers' gender observations. Through the implementation of a survey of research organizations that conduct telephone surveys, specifically, this paper addresses the following research questions:

1. How is respondent gender collected by survey research organizations?
2. What rationale(s) do organizations have for choosing a method to collect respondent gender?
3. How is information on respondent gender used by survey organizations (beyond inclusion in substantive analyses)?

The second paper explores covariates that may be predictors of observational gender classification error. A laboratory experiment tests causal hypotheses, which may explain gender misreporting. Expanding upon my preliminary research and guided by the linguistics literature, I test two situational predictors—placement and presence of noise in exposure—and discuss respondent and interviewer covariates of inaccurate judgments of respondent gender in telephone survey. Specifically, I address the following questions:

1. What is the relevance of two *situational* predictors of error in interviewer observations of a respondent's gender? Does allowing more time to disentangle gender cues improve observation and does a noisy phone room contribute to errors in observations?
2. What *respondent* characteristics, such as gender, race, and age, are covariates in interviewer misclassification of a respondent's gender?
3. What *interviewer* characteristics, such as gender and race, are covariates in interviewer misclassification of a respondent's gender?

The third paper evaluates the impact and potential consequences of errors in interviewer observations of respondents' gender on survey estimates. For example, how does President Obama's approval rating change among men when using the interviewer judgment of gender versus the true value, the respondent report? Analyzing pooled datasets collected by The Marist Poll, this final paper will address the following questions:

1. What differences in survey estimates are obtained and what is the bias when using interviewer observations of respondent gender for analysis?
2. Would different conclusions be made when using true-values of gender versus interviewer observations of respondent gender to identify statistical differences between male and female survey estimates?

1.1. Literature Review

Research from multiple disciplines contributes to the understanding of errors in aural observations. Within survey methodology, the paradata research provides guidance, while linguistics and psychology also offers information.

1.1.1 Linguistics

Strand and Johnson (1996), referencing other work (Peterson and Barney, 1952), note: "there is a lack of acoustic invariance in the speech signals as produced by different talkers" (p. 87). That is, there is significant variability in the speech properties across speakers. Linguistics research helps us understand these properties and documents many vocal cues and characteristics that contribute to a listener's ability to distinguish, and perhaps, confuse, a speaker's gender. Detailed in the following literature review from linguists, social psychologists, sociologists, as well as survey methodologists, differences between the vocal cues and characteristics of men and women can generally be classified into two types: physical characteristics and social conditioning.

Physical characteristics in vocal distinction of gender. As Owren et al. (2007) describe, the biological and anatomical differences between men and women provide the most basic and "stable" cues in discerning a talker's sex. Words or syllables spoken by a female talker will have different acoustic characteristics than if spoken by a male due to physiological differences.

A brief look at anatomy is helpful in understanding distinctions in voice. Graddol and Swann (1989) explain that when humans breathe out, the windpipes carry air from the lungs to the mouth cavities. The airflow process is uninterrupted—until one speaks, that is. When speaking, air first passes through the vocal folds (vocal cords) and act "like a pair of lips." A "coarse, buzzing sound" is created and fills the vocal cavities. The cavities then produce resonances, or formats, yielding the "speech-like quality to the noise that emerges from the mouth" (p. 14). Pitch is conveyed by: (1) the fundamental frequency (F0), defined as the rate of vibration of the vocal cords which is measured in

Hertz (Hz); and (2) the resonant structure which is the sound produced from the buzzing of the vocal cords. Both F0 and the formats are susceptible to the size, shape, and length of the vocal cavity. Laver and Trudgill (1979) observe that larger men normally have longer vocal tracts and vocal folds, which produce lower frequencies. In other words, a person's voice will usually reflect his or her physical structure.

Researchers have documented the physical differences between male and female voices. Vocal tracts are about 15% shorter in females (Goldstein, 1980), contributing to the likelihood that women will have higher and wider pitch ranges. Women's F0 is as much as 1.7 times those of men (Klatt and Klatt, 1990; Peterson and Barney, 1952). Specifically, Parris and Carey (1996) note that male speech pitch (F0) is typically between 60 and 120Hz, whereas female F0 is generally between 120 and 200Hz. However, other ranges are documented (e.g., Boone (1997) suggests that males have an average pitch of 120Hz; females average 220Hz). Regardless of the exact ranges, Peterson and Barney's (1952) laboratory study, in which 76 speakers recorded 10 monosyllabic words (e.g., *heed, hid, heard, had*), showed that males and females have some expected high and low ranges. The recordings were played over a high quality loud speaker in an auditorium and 70 listeners were asked to identify the spoken words. The researchers concluded, "in general, children's formats are highest in frequency, the women's intermediate, and the men's formats are the lowest in frequency" (p. 183).

Two more physical features of voice that differentiate male and female speakers are breathiness and articulation. Klatt and Klatt's (1990) analysis of two sentences spoken by 10 females and 6 males found females significantly more breathy than males. Breathiness, contributing to a "lighter" voice, is a product of the vocal cords not fully

closing when vibrating (Graddol and Swann, 1989), an occurrence more prevalent in females (Hanson, 1997). Another physical difference between male and female voices includes the tendency for females to speak with greater articulation (Traunmüller, 1997).

Social conditioning in vocal distinction of gender. In addition to basic anatomical differences, there are also more adaptive and socially conditioned features that differentiate the male and female voice. On average, women have a slower speaking rate (Picheny et al., 1985), tend to include more rising intonation (especially American women (Brend, 1975)), and are more breathy (Klatt & Klatt, 1990). Although breathiness is anatomically driven, Tuomi and Fisher (1979) found that women, when wanting to portray seductiveness, became increasingly breathy. In fact, both men and women were found to lower their F0 and speak slower when simulating a "sexy" voice.

Additionally, linguistics researchers suggest there are clues of a speaker's gender from spoken language and the content of speech. In a review of how female speech differs from male speech, Lakoff (1975) anecdotally identified several differences including: (1) women tend to use more "hedges" and qualifiers (e.g., *well, you know, I think, kinda*); (2) women are more likely to pose statements in a question form such as "*we're almost done with this questionnaire, aren't we*?" compared with men who are more likely to use declarative statements; (3) women employ words men typically do not use (e.g., *mauve*); (4) women tend to use more polite terms such as "*please*" and "*thank you;*" and, (5) women often come across as more compassionate, gracious, and expressive. In contrast, Latkoff notes: (1) men tend to be more direct; (2) men are more likely to discuss subjects such as sports, money, and business; and, (3) men typically make more quantifiable and concrete references to time, space, amount, and objects.

In her review of "informal observations, speculations, and stereotypes...[and] a report of empirical findings" Haas (1979) summarized:

> Women's speech is said to contain more euphemisms, politeness forms, apology, laughter, crying, and unfinished sentences. They are reputed to talk more about home and family and be more emotional and positively evaluative. Further, women's speech is stereotyped as nonassertive, tentative, and supportive. Women are also said to talk more than men.
>
> Men, on the other hand, are reputed to use more slang, profanity, and obscenity and to talk more about sports, money, and business. They are reputed to make more hostile judgments and to use language to lecture, argue, debate, assert, and command. (p. 623)

Consistent with these earlier observations, Newman et al. (2008) document how men and women use language differently. In their study, electronic text samples (both written and spoken), representing 70 studies from 22 laboratories (collected from 1980-2002) were coded through a text analysis program, Linguistic Inquiry and Word Count (LIWC). They found that women, when compared with men, use more negations, pronouns, verbs, "social words," "references to home," and "psychological process references" (p. 223). Men are more likely to use "a number of linguistic dimensions including word length, numbers, articles, and prepositions" as well as swear words. In terms of topics of conversation, men talked of "various current concerns" more than women. In general, findings "suggest that men, relative to women, tend to use language more for the instrumental purpose of conveying information; women are more likely to use verbal interaction for social purposes with verbal communication serving as an end in itself" (p. 212). Stereotypical male and female language patterns provide useful covariates when assessing interviewer judgments of a respondent's gender.

Survey interviewers can use these basic aural properties and subtle stereotypical language hints to aid gender assessments. They may also have direct language cues to

inform their observations of a respondent's gender. For example, telephone surveys will often incorporate some form of a household selection process that elicits gender references such as "*hold, on, let me get my husband on the phone"* or "*my wife has the next birthday; let me see if she's available.*" While standardized interviewing techniques typically attempts to minimize side dialogue, certain subjects or questions organically elicit more response, emotion, or reaction from male respondents than female ones or vice versa. For example, a questionnaire that is exploring shopping habits may invoke comments from male respondents such as "*I never buy anything, my wife does all of the shopping*."

Measurement error in gender assessments. Some research finds high accuracy in listeners' ability to make aural determinations of a speaker's gender. In a set of laboratory experiments to determine how listeners differentiate a speaker's gender, Graddol and Swann (1989) found a marked gap between the *average* male and female voice—even though pitch varies from person to person—making inferring a person's gender straightforward. They concluded that when listeners use expected average pitch ranges to discriminate gender, accuracy is typically high, with little error. In another study, Oksenberg et al. (1986) asked raters at the University of Michigan to code personal characteristics from recordings of experienced interviewers' survey introductions (approximately 30 seconds long). They found that higher pitched voices appear to be associated with females and lower pitches are expected for men. In fact, referencing other work (Smith, 1979; Pear, 1931; Harms, 1961), the authors motivate their research by stating: "Sex, age, social status, and race are accurately identified from the voice" (p. 99). Their findings infer that men and women are distinguishable.

However, other research documents notable measurement error in gender identification. Listeners can fail to make correct gender classifications due to overlaps in pitch (Hess, 1983; Shimamura and Kobayashi, 2001; Ross et al., 1974) or vocal properties (Graddol and Swann, 1989; Mendoza-Denton and Strand, 1998). Harb and Chen (2005) provided a summary of these findings, stating, "A good estimate of the pitch can only be obtained for voiced portions of a clean non-noisy signal. Moreover, an overlap of the pitch values between male and female voices naturally exists, hence intrinsically limiting the capacity of the pitch feature in the case of gender identification" (p. 3). The researchers found that non-ignorable overlap exists.

Research by Graddol and Swann (1989) also shows a lack of a clear delineation of vocal pitch. Their study charted the pitch ranges of 27 staff members (12 males and 15 females) at Open University and found sizeable variability in pitch within gender (Figure 1.01). Moreover, they document significant overlap in female and male pitch ranges (Figure 1.02). It is this overlapping portion of male and female vocal ranges that Oates and Dacakis (1983) described as *the gender ambiguous range*. They observe that although women and men tend to use the extremes of their pitch differently, they both share an extensive middle range.

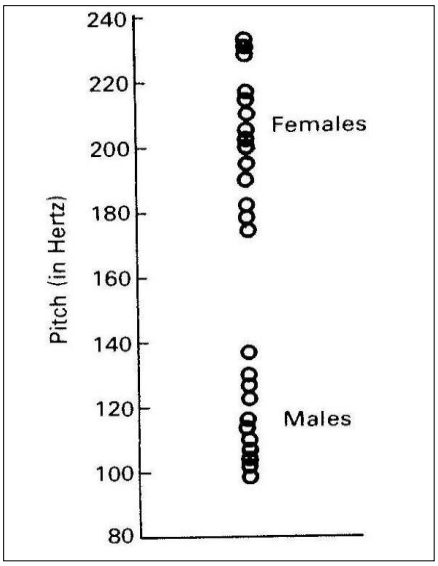


Figure 1.01: The Average Speaking Pitch of a Sample of Men and Women (Figure source: Graddol and Swann (1989, p. 20))

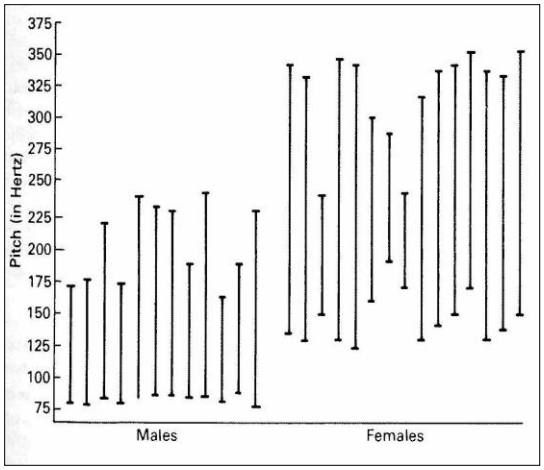


Figure 1.02: The Extent to which the Pitch Ranges of Men and Women Overlap (Figure source: Graddol and Swann (1989, p. 21))

Given that pitch ranges between female and male voices are non-exclusive, Strand and Johnson (1996) conclude they should be described along a continuum and not dichotomously as female vs. male. The concept of gender being classified along a spectrum of very masculine to very feminine is not new and has been documented (Coleman, 1976). For survey methodologists, overlapping patterns of speech between men and women point to the potential for at least some level of error in interviewer observations of respondent gender.

Identifying gender from voice is also compromised by an individual's ability to adapt, tailor, and manipulate their speaking style (Klatt and Klatt, 1990). In a study investigating the perceptual and acoustical components of voice, Andrews and Schmidt (1997) tasked 88 undergraduate student listeners to evaluate 22 voice samples provided by biologically male speakers who identified themselves as transsexuals. Speakers recorded a passage first using their masculine voice, then their feminine voice. Results

indicated that listeners could distinguish the masculine and feminine sounding voice samples. Femininity was associated with higher fundamental frequency, increased breathiness, and greater animation. Gender identification, especially among women speakers, is also affected by hormonal activity. Women taking hormone treatments typically experience a lowering of their vocal pitch (Damste, 1964); pregnancy as well as menstruation also alters pitch (Graddol and Swann, 1989). What emerges from this research is that individuals are capable of vocal accommodation making delineations between male and female voice less clear.

Behaviors, such as habitual smoking, excessive yelling or strain of the voice (e.g., vocal fatigue), and extreme alcohol use could change the expected properties of male and female voices (Graddol and Swann, 1989; Welham and Maclagan, 2003). Medical conditions (e.g., laryngitis or a common cold) (Graddol and Swann, 1989), or sleep deprivation (Bagnall et al., 2011) could also alter vocal norms.

The setting in which telephone interviewers are asked to make gender observations may contribute to an increase in errors beyond what has been documented by linguists. As Harb and Chen (2005) commented, telephone communication is often a noisy signal, especially in a telephone data collection facility. Centralized telephone data collection centers often contain many stations with interviewers sitting in close proximity. Auxiliary sounds and distractions may make detecting nuanced distinctions in vocal properties more difficult for interviewers. Because gender assessments are typically made quite early in the survey if used for screening and filtering, limited acoustic information is available for interviewers to disentangle distinctions in voice. Interviewers are forced to use impulse, expectations, and perhaps, stereotypes, to make gender

determinations. In addition, much of the research conducted in linguistics is in an experimental lab setting in which the observation and classification of the speaker's gender is the central task. For interviewers, this is only one of many tasks. Thus, the attention and thought applied to the task in a survey setting versus an experimental lab is considerably different.

Even small errors may have significant consequences on survey estimates. For instance, what are the implications for survey researchers, especially if certain interviewers have more errors in their judgment than others? As Singer et al. (1983) pointed out, bias—both response and interviewer—is a significantly greater threat in telephone surveys, because usually fewer interviewers complete more surveys and workload is higher when compared with other data collection modes.

The type of misclassification is also important to consider. When interviewer observations are used to determine survey screening and logic, random misclassification of gender, that is, equal misclassification of men and women, may disrupt the intended questionnaire flow, but not necessarily affect survey estimates. This could decrease the efficiency of nonresponse adjustments. However, if such misclassification is systematic (non-random), bias in final survey estimates could result.

Covariates of misclassification from linguistics. What influences the accuracy of interviewers' gender observations? As discussed, when interviewers hear a respondent's voice, that voice contains cues from a variety of anatomical and adaptive properties. Yet, those cues and therefore interviewers' judgments are also informed by other characteristics; specifically, covariates, such as the speakers gender itself, their race and age, the length of listening time, and the experience of the listener.

In addition to knowing the rate at which an interviewer can accurately classify a speaker's gender, one must also understand the predictors of the interviewers' observational errors. Useful covariates are documented in the linguistics literature. In fact, research has shown that the gender of the speaker is a significant covariate. Owren et al. (2007) found that listeners are able to determine the sex of males easier than females. In their discussion of talker-sex perception, referencing the marked fundamental frequency of vocal fold vibration and format frequency in males, they concluded that listeners have an advantage when perceiving the sex of males. They stated:

> Because sexual selection leads males to diverge from the 'default' female form [i.e., physiological changes in speech structures occurring at puberty], adult male voices can be considered 'marked' by the sexually selected features of lowered F0 and formant frequencies. It therefore follows that listeners should hear talker sex somewhat more easily in male than in female voiced sounds. Specifically, the presence of critical features of "maleness" [low F0, low formants] virtually guarantees that the talker is an adult male. However, their absence does not unequivocally imply that the talker is an adult female. (p. 930)

Using undergraduate students as participants, their series of lab experiments supported hypotheses that listeners would more accurately identify the gender of males over females. Across the 80 participants, 60 produced higher quality judgments for male vowel segments while only 13 performed better for females sounds (7 identified stimuli equally). Additionally Owren et al.'s research found that many of the listeners were able to classify male voices faster than females.

Other work has supported these findings. Coding gender of 10 male and 10 female student and faculty talkers at the University of Oregon in a laboratory setting, Coleman (1976) determined 88% of all speech samples (first 7 lines of the "Rainbow Passage" and 2 isolated vowel sounds) were correctly identified. However, differential

classification errors were observed. The 15 undergraduate coders accurately determined the gender of 98% of all males, while females were accurately identified 79% of the time. Likewise, Lass et al. (1976) documented that in listening to short vowel segments, 96% of voiced vowels were accurately classified male or female (although only 75% accuracy was obtained for whispered segments). His study found that females were more likely to be wrongly assigned (96.5%-100% accuracy for men compared with 87.0-97.5% for females). Although the focus of Loebach et al.'s (2009) study was to assess the implications of training of speech on the output of a cochlear implant, their research design included 24 normal hearing subjects tasked to transcribe sentences (e.g., "*The beauty of the view stunned the young boy*") in a laboratory setting from 20 male and 20 female speakers. They found that subjects were 77% accurate in their judgments of gender and that they were more correct in their classification of male talkers (83%) than females (71%).

In Honorof and Whalen's (2010) laboratory experiment, 12 participants (6 males; 6 females) listened to recordings of both males and females saying the vowel "ah." In coding the speakers' gender, the study found that, while males were more accurately coded at low frequencies and females identified more accurately at high frequencies, listeners were "especially inaccurate" for woman with low frequency voices. Moreover, coders were generally more confident in their identification of male voices. However, Krauss, Freyberg and Morsella (2002) suggested that female speakers were identified with "marginally better" accuracy than males; although differences in this study were not statistically significant (79% vs. 74.1%).

Race has also been documented as a valuable covariate (and found in preliminary research (McCulloch et al., 2010)). Lass et al. (1978) conducted a lab experiment with 20 participants (10 white and 10 black; 5 males and 5 females comprised each racial group) in which 30 white, female students observed the sex and race of the speaker. While the experiment varied the listening stimuli (tape recordings played forward, backward, and time compressed), their findings included the following: white speakers were more accurately identified than black speakers; observations of white males were of higher quality than those of black males and white females were more accurately categorized than black females. However, listeners in this study more accurately identified females than males overall. This may be due to the all-female nature of the evaluators. Lass et al. (1978) also investigated the confidence (on a scale of one to seven) of the evaluators' observations. They found that listeners were most confident in their judgments of female speakers, specifically white females over black females. However, their discussion did not include the relationship between confidence and quality of listener judgments.

In a separate study, Lass et al. (1979) reinvestigated the issue of speaker race and sex identification; this time, to understand the relative importance of phonetic complexity. Twenty speakers (10 white and 10 black; 5 males and 5 females in each group) each recorded 4 sentences. After listening to the recordings, raters (20 students at West Virginia University (10 males and 10 females) in a lab setting) indicated the sex and race of the speakers, as well as the confidence of their judgment. Lass et al. found, across all experimental conditions, listeners were 98.7% accurate in their sex observations of black speakers; 99.4% for white voices; 99.2% correct sex judgments of men; and 98.9%

accuracy among female speakers. While differences were relatively small, gender identification of white females was higher than black females.

Hughes and Rhodes (2010) documented a speaker's age as another predictor for identifying a speaker's gender from his or her voice. In their study, 97 raters coded both the age and gender of 101 recordings of individuals counting from 1 to 10. While they found widespread differences between the sex and age of the speaker, raters were the least accurate in their gender assessments of children and adolescents. Hughes and Rhodes note that this is not necessarily surprising given the similarity between male and female voices prior to puberty. Additionally, as they suggest, questions about gender identification remain, in part because research shows that as women age (especially after menopause) their voices typically deepen.

Language and ethnicity of a speaker have also been found to be a useful covariate in understanding gender identification from voice. Parris and Carey (1996) conducted a series of experiments using the British English gender identification system. Up to five seconds of speech in eleven languages were tested. Gender identification was found to be perfect among Mandarin and Vietnamese languages. However, other languages were inaccurate in 5.2% of cases (averaging 2.0% identification error overall). The average gender misclassification rate was 2%. Other research investigated pitch among speakers of various languages highlighting cultural differences (e.g., Loveday, 1981;). Majewski et al. (1972) found after listening to 103 male Polish university students and 157 American male students, that the mean F0 was significantly higher among Polish men when compared with American men (137.6 Hz vs. 118.9 Hz). These differences, along

with those for race of speaker, motivate the need to consider these covariates in understanding interviewers' gender judgments.

Most of the linguistics research cited is based on exposure to short vocal samples. Would an increase in the length of listening time improve observations and reduce errors? One study that examines this issue (Harb and Chen's, 2005) observed an improvement in overall gender classification accuracy with only small increases in exposure to voice samples. For one second of listening time, 93.5% of recorded telephone samples (nine male and nine female voices) were correctly identified as male or female. Yet, after five seconds of exposure, correct classification rates rose to 98.5%. These rates, however, only represent the outcomes of a single expert coder. Given that survey firms that use interviewer gender observations are likely to vary their placement in the questionnaire, this is an important covariate, which will be explored experimentally in my research.

Listeners themselves were also found to be a significant covariate in Nyggard and Queen's (2000) experiment to determine whether gender identification could be a learned task. They found that women had greater success in distinguishing both male and female speakers compared with men, while men identified female speakers with less accuracy.

Impact of stereotypes. Studies have shown how expectations do, in fact, alter conclusions. In a simple situation, listeners hear a higher pitched voice and conclude the respondent is a woman. Hawkins (1993) determined that listeners of vowel sounds produced by both black and white (both male and female) speakers associated lower F0 with African Americans; however, results across listener groups suggested that classification was based more on stereotypes, rather than physiology. Strand and Johnson (1996), Johnson et al. (1999) and Strand (2000) conducted a series of experiments to

understand how perceptions and expectations of a speaker's gender altered their identification of vowel sounds. Listeners heard a vocal segment presented with the visual cue of a male or female face (not necessarily the face or gender that recorded the vocal segment). They found that categorization of a component of the auditory signal could be changed simply by altering the gender of the face. Perceptions of the aural cue were affected by the gender of the person supposedly "producing" the sound. Thus, listeners' gender classifications were affected not only by what they saw, but what they imagined.

Evidence of stereotyping and its influence on gender classification was also explored in Imhof's (2010) experiment where 96 (48 males and 48 females) German university students listened to a set of voice recordings from both men and women. The recordings were meant to appeal to gender stereotypes such as "how to repair the inner tube of a bike" for men; "how to prepare a shortcake" for women; and reading directory names and addresses for a gender-neutral text. The recordings were modified to create both high and low pitch versions. After hearing each of the voice segments, coders documented the perceived gender, age, height, body type, personality type, and attractiveness of the talker. In terms of gender-related findings, Imhof found that female voices were assumed to be significantly younger when compared with male voices. In addition, when talking about a male topic, female voices were assumed to be younger in age than when the same female talker discussed baking or directory information. These findings suggest that several stimuli, including how and what is said during an interview, contributes to the determination of a speaker's gender.

1.1.2. Observation of Respondent Gender as Paradata

Whether they are intentional or involuntary, interviewers make judgments about their survey respondents (Hyman et al., 1954). In recent years, the use of interviewer observations has been expanded and is requested across various modes of data collection. Now interviewer observations are analyzed in the form of paradata—information about the process of survey data collection, such as time stamps and call records (Couper, 1998). Several prominent face-to-face surveys, such as the U.S. National Survey on Drug Use and Health, the U.S. Survey of Consumer Finances, the American National Election Study (ANES) and the European Social Survey (ESS), require interviewers to record their observations about neighborhood characteristics to guide survey design and nonresponse bias assessment (Kreuter and Casas-Cordero, 2010). In fact, the U.S. National Survey of Family Growth (NSFG) asks interviewers to make observations about the presence of children in the household and guess whether the respondent is currently in an active sexual relationship (Lepkowski et al., 2010; Groves et al., 2007; West, 2010a). Such interviewer observations are often used to guide fieldwork decisions and, in some instances, augment nonresponse adjustments or assess nonresponse bias (Copas and Farewell, 1998; Lynn, 2003; Maitland et al., 2009; Groves and Heeringa, 2006; Lepkowski et al., 2010; Kreuter et al., 2010).

An investigation into the formation and quality of telephone interviewers' observations of respondent qualities also requires an understanding of how judgments are created. Are interviewers able to form accurate judgments based solely on aural communication, or are non-verbal cues essential to accurate observations? In other

words, beyond the basic linguistic features, is asking telephone interviewers to determine a respondent's gender using only voice a reliable task?

Although paradata researchers have yet to provide evidence for the quality of interviewer observations, researchers in social psychology offer support for its application. They have shown that such impressions, even those based on gut reactions or intuition are informative and often accurate. In his dissertation proposal, West (2010b) discusses and summarizes the following concepts. Whether making and recording a doorstep (in-person) judgment or one determined solely using vocal cues, interviewers are making what is known as impulse, or "thin-slice" observations (Ambady et al., 1999). In fact, such "first impressions" obtained from brief exposure or behavioral observations are often correct (Winerman, 2005). These judgments, made from brief exposure, are encompassed in the "zero-acquaintance paradigm" (Albright et al., 1988), in which individuals make conclusions about the characteristics of strangers. Investigating this theory, Ambady et al. (1995) determined that female judges were more accurate in their "thin-slice" judgments—a notable covariate to consider when investigating the predictors and accuracy of interviewer paradata. West (2010b) also notes research, which found that the emotional state of the observer appears to be correlated with the accuracy of his or her observations. Ambady and Gray (2002) found judgments made by those feeling sad were less accurate. This is an interesting consideration for survey researchers whose interviewing staff may be comprised of individuals on any given day with a variety of emotional dispositions. While it is unlikely that the effect of interviewers' emotional states would result in systematic errors in gender judgments, it is, however, a conceivable

notion if certain interviewers are routinely sadder and disproportionately affecting misclassification.

Should certain interviewers be less adept at making accurate observations about a respondent, the utility of their paradata is compromised. Thus, the impact of any one interviewer in a telephone survey is magnified. While one study suggests that gender misclassification is clustered around specific interviewers (McCulloch et al., 2010), other research points out that rater agreement is typically high and reflects vocal stereotypes, not necessarily accurate judgments (Oksenberg et al., 1986). While high inter-rater agreement is, of course, possible, these examples from the literature raises the need for greater investigation of how observational errors cluster by interviewers.

Relatively little is known about the measurement error properties of survey interviewer observations. In a General Social Survey (GSS) Methodological Report, Smith (1997) evaluated the accuracy of observational paradata in a face-to-face setting - interviewer judgments of respondent race compared with true values (respondent self-identifications). Smith found an overall misclassification rate of between 3.3%-5.5%[4], concluding "minimal differences between racial classifications by interviewer observation and self-identification" (p. 4). The paper did not evaluate the possible consequences of these errors; rather only providing base rates of race misclassification.

[4] This range is due to the how the misclassification is calculated (e.g., treatment of missing data, mispunches, refusals, and differences in race terminologies). 3.3% misclassification was documented when only looking at cases in which interviewer observations of race were different from that self-identified from the respondent.

Beyond Smith's work, only recently have researchers started to thoroughly investigate the quality of interviewer observations and determine the extent to which inaccurate interviewer judgments lead to survey errors (Eckman, 2010; Sinibaldi, 2010; West, 2010a). Although interviewer observations show promising applications for nonresponse adjustments, Casas-Cordero (2010) documented sizable measurement error for in-person interviewer observations of area and household characteristics in the Los Angeles Family and Neighborhood Survey (LA FANS). Using the NSFG, ANES, and ESS, Kreuter et al. (2010) found that interviewer observations had only weak correlations with the response propensity. However, using such paradata for nonresponse adjustments led to changes in estimates only when the interviewer observation was highly correlated with the variable of interest. Thus far, the work in this area has been limited to face-to-face surveys. The only study to date addressing telephone interviews is McCulloch et al. (2010), which found 92% accuracy of interviewer observation of respondents' gender (detailed findings are discussed in the preliminary research section). The degree of measurement error of telephone interviewer observations is unknown.

In addition to interviewer observations, audio and visual recordings of the interactions between respondents and interviewers are another form of paradata that are useful in the assessment of measurement error. Kreuter and Casas-Cordero (2010) documented the limited strategies that telephone interviewers can employ to increase survey cooperation. Here, recordings provide paradata to inform both survey participation and question level measurement errors. Conrad et al. (2013) drew upon over 1300 audio-recordings of surveys interviews to identify speech behaviors that related to decisions to participate in the survey, refuse, or schedule a callback. The use of

these paradata in this research provides helpful information to inform interviewer training and data collection procedures. Also using recordings of telephone surveys, Jans (2010) analyzed vocal characteristics such as speech rate, interruptions, pitch tones, disfluencies, and pauses to successfully understand item nonresponse in income data.

For the assessment of gender observations, these paradata will be used in this research to (1) identify characteristics of respondents and interviewers that affect the accuracy of interviewer judgments; and to (2) assess the perceived sensitivity of asking a respondent their gender (e.g., hesitation, laughter, comments).

1.2. Preliminary Research

To present the need for this dissertation research, I conducted preliminary work to document error rates in interviewer observations of respondent gender. I first present findings from a study using a single data source. I then document misclassification rates from other data sources.

1.2.1. Gender Observation Quality in Single Data Sources

Using data from 28 independent public opinion telephone surveys conducted by The Marist College Institute for Public Opinion (MIPO), McCulloch et al. (2010) conducted a preliminary analysis of the quality of interviewer gender observations in telephone surveys.

Methods. All data were collected between September 2008 and February 2010 from their centralized telephone facility in Poughkeepsie, New York. Adult residents (aged 18 years and older) from three sample frames were represented in the data (United States, New York State, and New York City) with each study covering a variety of topics

such as politics, current events, and social attitudes. Only data collected from random-digit-dial (RDD) landline telephone frames were included; data from the cell phone frames were excluded. Across all surveys, 25,138 cases were available for analysis.

The majority of respondents were over the age of 45 and white. Gender proportions were more evenly distributed (44% male; 56% female). A total of 475 unique interviewers completed at least one survey included in our study. All MIPO interviewers were undergraduate students at Marist College and underwent extensive telephone survey training. Although interviewers were not specifically trained in distinguishing gender voices and pitch tones, they were told of the importance of their gender assessment. By virtue of being undergraduate students, all interviewers were between the ages of 18 and 23.

Aside from containing respondent and interviewer characteristics, the unique feature of this data set was that all cases included both an interviewer observation of respondents' gender (*observe*) as well as a respondent self-reported gender (*reported*). The placement of both items was consistent across all studies. MIPO interviewers were asked to record the respondent's gender immediately after the survey introduction likely having heard only a few words from the respondent. Respondents were also asked to state their gender at the very end of the questionnaire.

Results. Overall, assuming the respondent reported gender is the true value, 8.3% of the 25,138 judgments made by the interviewers were incorrect. As shown in Table 1.01, there was significant differential measurement error between male and female respondents. In fact, 12.6% of all female respondents (n=14,194) were incorrectly

observed male by interviewers—a sharp contrast to the 2.6% of male respondents (n=10,944) who were misclassified.

Table 1.01

Respondent Reported Gender vs. Interviewer Observed Gender

MIPO Data	Respondent reported male	Respondent reported female	Total
Observed Male	97.2%	12.6%	49.5%
Observed Female	2.6%	86.9%	50.2%
Cannot Make a Guess	0.2%	0.5%	0.3%
Total	100%	100%	n=25,138

Interviewer gender observations revealed some systematic differences across various racial groups of respondents.[5] African-American respondents were more likely to be miscoded than other respondents. Interviewers incorrectly coded nearly 13% of all respondents in this subgroup. Mirroring the overall findings, 18.1% of African-American women were miscoded as men, while only 2.9% of African-American men were perceived to be women by interviewers.

Multivariate regression models were used to determine whether interviewer or respondent characteristics could predict mismatches between interviewer observations and respondent reports of gender. Accounting for the clustering of respondents by interviewers via the inclusion of random interviewer effects, the hierarchical linear

5. Other demographic subgroups were analyzed in this way but did not show notable differences.

probability models (LPM) used an error between interviewer observation and respondent report (where 0 = no error and 1 = error) as the dependent variable. Given the inclusion of interaction terms in the nonlinear model, an LPM was fit to ease interpretation (Mood, 2010). Various respondent demographics, interviewer covariates, and controls for potential confounders such as household selection were added as predictors. In Table 1.02, two LPMs were estimated to determine the probability of making an incorrect gender observation as a function of various respondent and interviewer characteristics. Model one includes only main effects while model two includes respondent and interviewer interaction terms. Asterisks denote some significant indicators of the probability of making an observational error, especially by gender and race. For example, looking at the respondent's race in model one, the African-American coefficient indicates that when the respondent is African-American, it increases the probability of making an incorrect gender judgment by 2.4%. Older respondents are less likely to be miscoded, and as previously noted; women are more likely than men to be misclassified by interviewers.

Looking at interviewer characteristics, experienced interviewers (measured by their total number of completed telephone interviews at MIPO) were more likely to make observational errors than those with less interviewing experience. A possible explanation for this finding is that experienced interviewers, not in a supervisory role, are moving quickly through the gender observation task and perhaps more focused on obtaining completed interviews than the quality of their gender judgment. Less seasoned interviewers (who have been more recently trained and told of the importance of their gender observation) may be spending more time on this item, thus increasing their

accuracy. Additionally, as indicated by high intra-cluster correlations (estimated ρ value =.096), there does appear to be substantial within-interviewer correlations in errors. That is, some interviewers were more likely to make gender observation errors than others.

MIPO selects household respondents by asking to speak with the youngest male currently at home. In cases where a male was asked to come to the phone, we assume the interviewer would already know their gender. To account for this, household selection was used as a control variable in the model. While knowing that a young male was asked to come to the phone reduced the probability of making a gender judgment error by 34.3%, it had little effect on the other coefficients.

Table 1.02

Hierarchical Linear Probability Model Results of Interviewer Gender Observation Errors

		Model 1	Model 2
Respondent Race (white omitted)			
Respondent Race African-American		0.024***	0.041***
		(0.00)	(0.01)
Respondent Race Hispanic		0.002	
		(0.01)	
Respondent Race Asian		0.023*	
		(0.01)	
Respondent Standardized Age		-0.006***	0.004**
		(0.00)	(0.00)
Respondent Reported Gender (males omitted)		0.336***	0.080***
		(0.00)	(0.01)
Interviewer Gender (males omitted)		0.006	-0.008
		(0.01)	(0.01)
Interviewer Race (white omitted)			
Interviewer Race African-American		0.012	0.002
		(0.01)	(0.02)
Interviewer Race Hispanic		-0.002	
		(0.02)	
Interviewer Race Asian		0.025	
		(0.04)	
Interviewer Supervisor (non-supervisors omitted)		0.010	0.002
		(0.01)	(0.01)
Experience			
Level 2 Interviewer (level omitted)		0.015	0.019
		(0.01)	(0.01)
Level 3 Interviewer		0.032**	0.040**
		(0.01)	(0.01)
Level 4 Interviewer		0.044**	0.038*
		(0.01)	(0.02)
Single person HH or youngest male		0.050***	
		(0.01)	
Youngest male comes to the phone		-0.343***	
		(0.00)	
Interaction: Respondent & Interviewer African-American			0.039
			(0.02)
Interaction: Respondent African-American/Interviewer			0.101*
Asian			(0.05)
			(0.02)
Interaction: Respondent &Interviewer Female			0.019**
			(0.01)
Constant		-0.348***	0.006
		(0.02)	(0.01)
Observations		24,188	24,188
Sigma_u		0.072	0.084
Sigma_e		0.221	0.248
Rho		0.096	0.104

*p<0.05, **p<0.01, ***p<0.0001

Model two included the interactions between respondent and interviewer characteristics. While there are not notable main effects for interviewer gender and race, some interactions were significant. Women were more likely to be miscoded by female interviewers than male interviewers. African-Americans were more likely to be miscoded than non-blacks when being interviewed by a non-black interviewer. Other respondent and interviewer race pairs were included in the model but did not show significant interactions.

In none of the surveys was the flow of the questionnaire dependent on the gender observation. Thus, we could not investigate the consequences of filtering or screening on measurement error. However, we were able to investigate potential effects of interviewers' misclassifications of respondents' gender if used for nonresponse weighting adjustments. Estimates shifted only slightly and weighting may not be dramatically affected if gender is not highly correlated with the variable of interest.

Of course, there were shortcomings and limitations of the research. We could not be sure that our treatment of the respondent reports as a true-value of gender were, in fact, true scores as no known research has investigated deliberate misreporting of gender. In addition, one survey firm with student interviewers collected all data used in this study; thus, we were not able to eliminate the possibility that findings were due to a "house effect." However, two studies conducted by two different firms (both commissioned by JPSM, The Joint Program in Survey Methodology at the University of Maryland), also showed measurement error in gender assessments, although the magnitude varied. In the first study of 1,653 United States adults, trained interviewers observed respondents' gender immediately after the survey introduction. As a final

questionnaire item, respondents stated their gender. Reflecting the MIPO findings, there was an overall error rate of approximately 9%. Interviewers miscoded 14.7% of female respondents and 5.3% of male respondents. However, findings in the second JPSM telephone survey of 1,569 United States adults differed. In this study, only 2% of all interviewer observations were incorrect and the direction of incorrect gender observations was not significantly different. Although the placement of the gender observation was the same in both the MIPO and first JPSM study, the respondent-provided report was extremely early in the questionnaire in the second JPSM study. My research seeks to address such discrepancies by determining the predictors of observational error.

Investigation of these data suggests small, though possibly important, differences in survey estimates when using the interviewer gender judgment versus the respondent report to analyze data. A secondary analysis of one of the 28 datasets provides some evidence. The study, conducted in September 2008, interviewed 851 adult Americans from a centralized telephone facility in Poughkeepsie, New York. Treating the respondent report as the true value, overall, 7.9% of interviewers' gender observations (asked early in the questionnaire) were incorrect in this study. Included in the survey, mostly of a political nature, was a question that asked, "*If the 2008 presidential election were held today, whom would you support if the candidates are John McCain, the Republican candidate or Barack Obama, the Democratic candidate?"* As shown in Table 1.03, when comparing support for presidential candidates between estimates derived from the interviewer observation of the respondents' gender versus the respondent report, deviations do exist. Although these differences are not significant in a t-test comparing the means, the observed changes do motivate a need to explore other

variables where greater gender differences may exist and/or the gender misclassification rate is higher.

Table 1.03

September 2008 Presidential Toss-up When Using Reported vs. Observed Gender (unweighted data)

	Respondent **reported** male	Interviewer **observed** male	Respondent **reported** female	Interviewer observed female	Total
John McCain	50.13%	48.93%	43.75%	45.07%	46.65%
Barack Obama	40.83%	42.25%	46.12%	44.86%	43.71%
Independent Candidates	1.03%	1.07%	.43%	.42%	.71%
Other	.78%	.80%	.86%	.84%	.82%
Undecided	7.24%	6.95%	8.84%	8.81%	8.11%
Total	100%	100%	100%	100%	n=851

1.2.2. Gender Observation Quality in Other Data Sources

Extending McCulloch et al.'s (2010) work and adding empirical evidence to document the magnitude of errors stemming from interviewer judgments of respondent gender, comparable data from three different survey organizations were used to answer the question: *What is the misclassification rate in interviewer gender observations across other organizations?*

Methods. To obtain additional estimates for the amount of error associated with interviewer observations of respondents' gender three separate random-digit-dial telephone studies were used. A different data collection firm conducted each study. The purpose of comparing data from various organizations was to determine whether a "house effect"—differences in survey estimates obtained between organizations (Smith, 1978)—

contributed to the errors observed by McCulloch, et al. (2010). To address overflow data collection needs, The Marist College Institute for Public Opinion (MIPO) commissioned three organizations to conduct data collection. The availability of data from three, separate firms enabled this "house effect" research. Each study used identical methodologies to collect both the observational and reported data. The samples were drawn from the population of individuals 18 years of age and older living in private households in the contiguous United States.

Each firm utilized its own in-house interviewer training, which for all of these organizations does not include specific training on distinguishing gender voices and pitch tones. Given that each firm fielded a questionnaire designed by MIPO, variability among the questionnaires was only in the core survey topical questions, not the screening or items used to determine gender. The placement of both the interviewer observation and respondent report was consistent across all studies. Interviewers observed and recorded the respondent's gender after the survey introduction and asking two questions ((1) Are you 18 years of age or older? (2) How many adults, aged 18 or older, currently live in your household?). Respondents were also asked to state their gender at the very end of the questionnaire. MIPO's procedure for household selection was to ask to speak with the youngest male at home if more than one adult lives in the household. If the respondent changed, interviewers repeated the gender observation once the respondent came to the phone. The survey introduction and questionnaire flow for each of the studies was:

> **Q1**. Hello. My name is <name>. I'm calling from Marist College. We're talking to people in your community and collecting opinions about issues facing residents. Are you 18 years of age or older?
>
> Yes

No – *Interviewer ask to speak to another member of the household who is 18 and restate introduction*

Q2. How many adults, aged 18 or older, currently live in your household?
1
2
3 or more

Q3. *Interviewer*: Record gender of person on the phone by observation only:
Male
Female

ASK IF Q2>1
Q4. May I please speak with the youngest male who is currently at home?
Yes – *Once new respondent comes to the phone, interviewer re-states survey introduction, makes another observation of respondent gender, then continues with full survey*
No – *Continue to full survey with person on the phone*

<SURVEY CONTENT>

Last Q. Are you:
Male
Female

The Marist College Institute for Public Opinion collected the data for the first dataset. A total of 1,235 adults, 18 years of age and older, residing in the continental United States were interviewed by telephone from September 26, 2011, through September 29, 2011 from MIPO's centralized telephone facility in Poughkeepsie, New York. The topic of the survey was impressions of the elderly and expectations for aging. A total of 217 interviewers worked on this study. All MIPO interviewers were undergraduate students at Marist College and underwent telephone survey training. Approximately 20% of the completed interviewers were from a cell phone sampling frame, 80% from a landline frame.

Luce Research, located in Colorado Springs, Colorado collected the data for the second survey. A total of 1,042 United States adult residents (aged 18 and older) were interviewed by telephone between September 13th and September 14th 2011. A total of 132 interviewers completed interviews on this study. Demographic characteristics of the

interviewers were not available. The content of the questionnaire included current political and economic issues such as approval of President Obama, voter preferences for the 2012 presidential election, and opinions of various government entities and leaders. Approximately 20% of the completed interviewers were from a cell phone sampling frame, 80% from a landline frame.

For the third and final dataset, Opinion Search—a large data collection contractor—conducted 1,018 telephone interviews from January 6, 2011 through January 10, 2011. Calls were conducted from their Ottawa, Canada and Chicago, Illinois facilities. Exactly 100 interviewers completed interviews on this study. Interviewer demographics were not available. The questionnaire included a range of topics such as approval of President Obama, impression of John F. Kennedy, and opinions about the economic recession.

Results. Overall, as shown in Table 1.04, there are similar error rates in interviewer's observations of respondent gender across the three datasets analyzed. Firm 1 obtained a gender misclassification rate of 4.29; Firm 2, a rate of 4.42; and Firm 3, a rate of 3.34. In order to statistically test the difference between the mean error rates of these three separate samples, an analysis of variance test was used. Here, there is confirmation that significant "house effects" are not observed with an overall F value of 1.31 (p=0.269).

Table 1.04

Overall Respondent Gender Misclassification Rate for Three Data Collection Firms

	Overall proportion of misclassified cases	Number of cases
Firm 1	4.29%	1234
Firm 2	4.42%	1036
Firm 3	3.34%	1018

Confirming previous research, differential measurement error between men and women were observed across each data source. Presented in Table 1.05, women were misclassified at a significantly higher rate than men. McCulloch et al. (2010) found that women were 4.3 times more likely to be miscoded than men by interviewers. For Firm 1, women were nearly four times more likely than men to be misclassified; 1.65% of all male respondents were incorrectly judged while 6.38% of females were wrongly observed to be men. However, the differential misclassification rates obtained for Firms 2 and 3 are slightly smaller. Here, women were approximately 2.5 times more likely to be miscoded male. Chapter 3 of this dissertation explores the covariates of misclassification.

Table 1.05

Respondent Misclassification Rate by Gender for Three Data Collection Firms

	Proportion of misclassified male cases	Proportion of misclassified female cases	Number of cases
Firm 1	1.65%	6.38%	1234
Firm 2	2.61%	6.40%	1036
Firm 3	1.87%	4.66%	1018

The error rates observed in this research are notably lower, about half, than those documented in McCulloch et al.'s (2010) study which found an overall gender misclassification rate of 8%. There is, however, one key difference in methodology between the studies. In the McCulloch et al. (2010) data, interviewers observe respondent gender *immediately* after asking whether the respondent was 18 years of age or older. In the data from the three other firms, interviewers observed respondent gender after hearing the response to one additional question (*How many adults, aged 18 or older, currently live in your household?*). Although the verbal cues obtained by a response to a single additional question may appear inconsequential, research does suggest that

accuracy improves with exposure. Harb and Chen (2005) observed an improvement in overall gender classification accuracy with only small increases in exposure to voice samples. For one second of listening time, 93.5% of recorded telephone samples (nine male and nine female voices) were correctly identified as male or female. After five seconds of exposure, accuracy rose to 98.5%. Chapter 3 of this dissertation will experimentally determine whether small increases in exposure do, in fact, increase interviewer accuracy.

1.3. Dissertation Papers

My dissertation is structured in three journal-style papers. The first paper (Chapter 2) documents the common practices in collecting respondents' gender in telephone surveys across the industry. The second paper (Chapter 3) experimentally tests causal hypotheses of error in gender observations. The third paper (Chapter 4) examines and discusses the consequences to estimates of errors in interviewers' observations of respondents' gender.

1.3.1. Paper 1: Determining Respondents' Gender in Telephone Surveys: How Hard Can it Really Be?

The first paper of my dissertation identifies common practices in the collection of respondents' gender in telephone surveys. The purpose is to document how gender is collected and the ways which firms use interviewers' gender observations. Drawing a census of firms from multiple sampling frames, I conduct a primary data collection through a survey of research organizations that perform telephone surveys (289 firms participated). The resulting data informed the following research questions:

1. What are the different methods used by survey research organizations to collect respondent gender (including training practices), and how many firms use each method?
2. At what point in the questionnaire do firms collect respondent gender data?
3. What rationale(s) do organizations have for choosing a method to collect respondent gender?
4. How is information on respondent gender used by survey organizations (beyond inclusion in substantive analyses)?

Documenting the prevalence of the use of interviewer observations to collect respondent gender, the results reinforce the need for this dissertation. In addition, survey findings related to where interviewers are instructed to observe respondent gender and the perceived quality of judgments provided a foundation for subsequent chapters.

1.3.2. Paper 2: Sources of Inaccuracy in Interviewers' Observations of Respondents' Gender and Its Impact on Nonsampling Errors

The second paper of my dissertation identifies and experimentally evaluates sources of error in interviewers' observations of respondents' gender—especially those correlated with the placement of the gender observation. Expanding upon my preliminary research and guided by the linguistics literature, I seek to determine the predictors and correlates of inaccurate judgments.

To test the effect of listening time and placement of the interviewer observation on quality of the judgment, I utilize recordings and vary the length of exposure to determine if, in fact, quality improves with increased listening. I speculate that many firms that collect gender by observation do so at the end of the questionnaire, allowing for longer

exposure and vocal cues to inform their gender judgment; this should be confirmed in paper 1. However, little research exists documenting the quality of such paradata and whether placement is correlated with accuracy. This study is different from much of the current linguistics research due to: (1) the noisy signal inherent in a centralized phone room; and (2) the non-salience of the gender judgment. For the former situation, linguists often use non-noisy, laboratory settings for most of their research, which is rarely the situation in which interviewers make gender observations. In the latter situation, the focused attention to recording gender is likely to make listeners more conscious of their judgment. This is unlike a normal survey setting where gender is one of many collected data points and interviewers accomplish this quickly to move to the next question.

This portion of my research is designed to address the following research questions:

1. Do situational circumstances, exposure length and noise, predict error in interviewer observations of a respondent's gender? In other words, does allowing more time to disentangle gender cues improve observation and does a noisy phone room contribute to errors in observations?
2. Are rater (survey interviewers*)* characteristics, such as gender, race and experience, significant covariates when assessing interviewer observations of a respondent's gender in various situational conditions?
3. What respondent characteristics, such as gender, race, and age, are important covariates of error in interviewer observations of a respondent's gender in the situational conditions?

4. What are the interaction effects of the situational, respondent, and interviewer, level predictors on misclassification error?

1.3.3. Paper 3: Understanding the Consequences of Observational Gender Error on Survey Estimates

The third paper of my dissertation relates the findings from Chapters 2 and 3 to identify the consequences of inaccurate interviewer observations of respondents' gender. As discussed in the preliminary research, there is some evidence that errors in interviewer gender observations change final survey estimates when analyzed by males and females. The chapter addresses:

1. What differences in survey estimates are obtained and what is the bias when using interviewer observations of respondent gender for analysis?
2. Would different conclusions be made when using true-values of gender versus interviewer observations of respondent gender to identify statistical differences between male and female survey estimates?

Using data from 28 independent public opinion telephone surveys collected by The Marist College Institute for Public Opinion (MIPO), I conducted secondary data analysis to evaluate measurement error induced by inaccurate interviewer observations of respondents' gender, and estimate the impact of this error across 50 survey results (inclusive of attitudinal, behavioral, and demographic outcomes).

2. Documenting Common Practices in the Collection and Uses of Respondents' Gender in Telephone Surveys

Gender is an important variable for survey practitioners as well as users of survey data. Differences between men and women are often reported by journalists seeking to explain such things as voting behavior; by researchers documenting employment statistics among other things; and by academics looking to better understand patterns of opinion, for instance. For survey data collectors, depending upon the goals of the research, respondent gender may be used to: (1) screen for study eligibility; (2) determine skip patterns; (3) foster, perhaps unintentionally, interviewer tailoring strategies;[6] and (4) contribute to nonresponse bias assessment (e.g., informing post-stratification weighting). Yet, with this fundamental role in both practice and research, survey documentation shows there is significant variation in how this variable is captured across survey organizations as well as from study to study within an organization.

Looking back to the field's original modes of data collection, respondent gender was, and perhaps still is, a seemingly straightforward measure. For face-to-face surveys, interviewers have visual cues leaving less of a reason to ask respondents their gender (although gender ambiguity is still possible). For mail surveys, the opposite is true—

[6] Although there is no evidence in the literature that interviewers employ different tailoring techniques based upon observed respondent gender, it is known that interviewers use different strategies informed by their ability, expectations, and impressions. Groves and McGonagle (2001) state: "experienced interviewers often report that they adapt their behavior to the perceived features of the sample unit" (pg. 250).

asking respondents for their gender was the only way to obtain the information. However, with the emergence of telephone surveys, the best approach to collecting respondent gender is less clear. Should interviewers ask whether they are speaking with a male or female or should they determine the respondent gender based on aural cues? Telephone data collection firms may have made the decision regarding how to determine gender using inferences and adopted practices from other data collection modes. Perhaps organizations assumed gender was as easy to collect by phone as it was for face-to-face interviews. Survey researchers did not standardize nor document 'best practices' in respondent gender collection methods. In fact, apart from looking at study documentation and survey questionnaires one by one, little is currently known about the ways in which survey firms collect and use respondent gender. Is there variation in how organizations approach the collection of this important variable? If so, what methods are used and how and why were they chosen? This research attempts to document these practices.

Why is it important to identify the prevalence and utility of practices in the collection of respondents' gender in telephone surveys? The answer comes down to potential problems and errors induced by various methods. If nearly all firms *ask* respondents whether they are a male or female, then error in gender data is likely small (assuming respondents are providing a true value). However, as suggested by looking at large-scale national studies cited in Chapter 1 (e.g., HINTS and BRFSS), some firms rely, at least in part, on paradata (interviewer observations) to collect respondent gender. Methods utilizing observational data may induce errors since interviewers could be wrong in their gender assessments. These errors may in turn affect quotas, survey routing

and logic (if gender is used as a filter or screen), and post-stratification adjustments (if used as a weighting parameter). Aside from logic and weighting implications, survey estimates could be affected if inaccurate gender data is used for any type of analysis (e.g., to discuss differences in opinions between men and women). In addition, at what point in a questionnaire gender is collected needs to be established. If firms observe gender early in the survey, presumably interviewers have little time to hear respondents' voices and distinguish confusing pitch signals. The decision of placement might then affect error rates.

Research documented in Chapter 1 of this dissertation shows some level of error in telephone interviewer observations of respondent gender. However, are these errors necessarily problematic for practitioners? If it is common practice for firms to collect gender strictly by asking respondents then there may be little reason for concern, as the potential problem is restricted to a small segment of the industry. However, if interviewers are routinely asked to judge whether they are speaking to a male or female by observation—and to do so early in the questionnaire with limited acoustic cues—there is reason for further exploration of the techniques and possible induced error.

The purpose of this chapter is to document how gender is collected and the ways in which firms use interviewers' gender observations (e.g., for survey logic/routing, for weighting, for analysis, etc.). Included here is an evaluation of the perceived accuracy of these judgments. Is it consistent with documented error rates? Through the implementation of a survey of research organizations that conduct telephone surveys, this chapter will address the following research questions:

1. What are the different methods used by survey research organizations to collect respondent gender (including training practices), and how many firms use each method?
2. At what point in the questionnaire do firms collect respondent gender data?
3. What rationale(s) do organizations have for choosing a method to collect respondent gender?
4. How is information on respondent gender used by survey organizations (beyond inclusion in substantive analyses)?

2.1 Data and Methods

Addressing these research questions requires primary data collection. To collect data from survey firms, an online survey of data collection organizations was implemented, with the goal of documenting practices in the collection of respondent gender. This section details the research methodology.

2.1.1. Frame Construction

Six frames, chosen to be as inclusive as possible of the various sectors encompassed by the field (e.g., marketing, academic, commercial, etc.), were used to construct the sample: (1) the American Association for Public Opinion Research (AAPOR) 2009-2010 Blue Book; (2) the 2010 List of Academic Survey Research Organizations (AASRO) published by the Survey Research Laboratory at the University of Illinois; (3) the 2010 membership list of the National Network of State Polls (NNSP); (4) the listings under the headings *Polling-Survey Research and Analysis* published by "Campaigns and Elections" magazine (C&E) in 2010; (5) the 2010 membership list of

the Council of American Survey Research Organizations (CASRO); and (6) the entries in the 2010 American Marketing Association (AMA) Green Book. The AMA frame included only those organizations falling under relevant survey research headings (e.g., *Interviewing Method: Telephone, Data Collection Field Services)*. The frames were de-duplicated and edited to include only unique firms that conduct survey data collection and are all located within the United States.[7] All 712 unique survey organizations, a census, were selected to participate in the study (Table 2.01).

Contact information (name, telephone number, and email address) was provided for most of the primary contacts in the frames. Especially for large firms, the primary contacts were in a marketing or sales capacity rather than operational personnel with knowledge of the targeted research questions. Well-documented hurdles of establishment surveys include low response rates (Rosen et al., 1991) and correct selection of the most appropriate respondents (Lynn and Sala, 2004). For firms where there was a known relevant contact person, the survey invitation was sent to that person. For firms that either did not have a contact name or email address listed or one whose title was clearly in a sales or marketing position, searches of the organizations' websites or calls to the

[7] Each firm was only included once in the study. Thus, using my own judgment and, when needed, information from the firms' websites, I de-duplicated the frames such that an organization (e.g., Marist College) which appeared in more than one frame (e.g., AASRO and AAPOR), was included only once. Study findings are not impacted by my subjective decisions concerning final frame assignments for duplicate cases. Only response rates (never the survey findings) are presented by frame. De-duplication was for the sole purpose of preventing over-coverage.

firms produced a more appropriate contact. However, if needed, the survey invitation asked invitees to forward the survey link to another person in their organization if he or she was better suited to complete the questionnaire.

2.1.2. Data Collection and Analysis

The questionnaire (see Appendix A) explored various aspects of gender data collection including: (1) collection method for obtaining a respondent's gender (i.e., whether the organization recorded by observation only, respondent report only, both, or neither (including placement, question wording, interviewer instructions, etc.)); (2) use(s) of gender data; (3) interviewer training related to the collection of demographic information; (4) interviewer monitoring and recording practices; (5) rationale for the chosen method for collecting gender; and (6) perceptions about the accuracy of interviewer gender observations.

In addition, organizational information such as age and industry sector were collected and used as covariates to determine whether practices are prevalent industry-wide or only among certain sectors. The questionnaire asked firms to indicate whether they are: (1) commercial, non-marketing (30% of the sample characterized themselves this way); (2) commercial, marketing (50% of the sample); (3) academic (15% of the sample); or (4) not-for-profit or something else (4% of the sample).[8] Organizations also provided their founding year. These self-reports of sector and age were used in the analysis to determine sector differences. A 1996 JPSM survey investigating the ethics of

[8] These percentages, and others presented in this chapter, may not add to 100% due to rounding.

survey organizations found that practices differed by age and sector of the organization (McCulloch and Presser, 2006). For example, small data collection firms were more likely to follow recommended standards for disclosure, and implement more rigorous interviewer monitoring practices. Given such differences, I expect to find significant variation in how gender is collected and differences by sector and age.

Data collection commenced on September 21, 2011. The survey was programmed and hosted online by Qualtrics Survey Software. Potential participants received an email invitation (see Appendix B) to provide the required components for informed consent (e.g., purpose of research, survey sponsor, administration length, confidentiality statement). A single question appeared on each online page; respondents advanced to the next question after entering their response. Appendix A shows both the questionnaire wording and some design features of the web survey (e.g., use of radio buttons, grid formatting for some questions, etc.).

Overall, 347 firms logged into and participated in the survey (281 completed and 66 partially-completed interviews); 51 did not conduct telephone surveys and another seven did not collect gender data. Including partials, this left 289 cases available for analysis. Treating the sample members that did not log into the survey as a refusal and using the AAPOR response rate calculator #4, the overall response rate was 48.7%. As shown in Table 2.01, response rates varied significantly by sampling frame. Partially completed cases were retained for analysis as long as data was obtained for the question asking how firms collect respondent gender. The field period remained open for a total of seven weeks. For four weeks, weekly email reminders were sent to nonrespondents to

reinforce the importance and legitimacy of the research, confidentiality of responses, and length of time to complete the survey (less than 10 minutes).

Table 2.01

Response Rate by Sampling Frame

		Number of Completed Interviews		
Frame	**Number of Frame Members**	**Full**	**Partial**	**Response Rate (AAPOR #4)**
AASRO	50	32	7	78.0%
AAPOR	60	30	5	58.3%
CASRO	273	119	16	49.5%
MRA	281	82	33	40.9%
C&E	32	11	2	40.6%
NCPP	16	7	3	62.5%
TOTAL	712	281	66	48.7%

Paxson, et al. (1995) found that telephone follow-ups in establishment surveys produced higher response rates. Within the two largest sampling frame—CASRO and MRA—approximately 10% of firms that did not respond after five weeks of fielding time were randomly selected for a telephone follow-up attempt. These two frames were chosen due to their large size and relatively low response rate. The questionnaire was reduced to four key questions (denoted by an asterisk in Appendix A). The 33 randomly selected nonrespondents (20 from the MRA frame; 13 from the CASRO frame) first received a final email invitation notifying that they had been chosen for an additional telephone follow up. At that time, six then completed the survey online. Using an internet search, names and phone numbers were matched to 25 of the 27 cases (a name or telephone number could not be matched to two of the email addresses). I conducted five interviews from the MRA frame and four from the CASRO frame via telephone. I left

messages with the remaining nonrespondents. The completed telephone interviews are treated as partial interviews and included in the rates presented in Table 2.01.

Item non-response was observed in 50 cases for the question that asked for the organization's industrial sector. Business email addresses were linked to 34 of these cases and values were imputed after looking at the firm's website describing the type of services they offer. Likewise, the year that the organization was founded was missing for 51 cases; 32 of which were then determined and imputed via verification from the organizations' websites. The survey did not allow for item nonresponse for many of the substantive questions, thus, item nonresponse was only obtained in the case of partially-completed cases.[9]

Although some may argue that significance testing is not necessary in a census because there is no sampling error (e.g., Selvin, 1957), others maintain statistical tests are still useful given that a census can be conceptualized as a sample in time or a sample from a larger, possible universe (Hagood & Price, 1952). Taking a more conservative approach, significant differences between means to identify demographic differences (e.g., age of organization, industry sector) are calculated using independent t-tests. For example, in Table 2.02, the asterisks denote that there are statistically significant differences between sectors of the industry at the 95th confidence level. Data were analyzed using both STATA 11 and SPSS 18.

---

[9] One organization expressed frustration with not being allowed to skip questions. Thus, a different module, allowing for item nonresponse, was sent to this firm.

2.2. Results

The findings from the survey of data collection organizations are presented in four sections. The first section quantifies the prevalence of gender collection methods; the second documents at what point in a survey firms collect gender; the third explains the rationale for the chosen approaches; and the fourth describes the utility of gender data from various respondent gender collection methods.

2.2.1. How is Respondent Gender Collected?

While nearly all (98%) organizations that conduct telephone surveys collect respondent gender, there is significant variation in how it is measured across firms. Variations of collecting respondent gender include: (1) asking the respondent (e.g., *Are you male or female?*); (2) interviewer observation only (e.g., *Interviewer: Record the gender of the respondent*.); (3) a hybrid of both observation and respondent report, called *ask-assisted* throughout this dissertation, where firms instruct interviewers to ask respondents their gender only when necessary or not obvious (e.g., *Interviewer: Record the gender of the respondent. Ask only if not obvious*.); (4) utilizing record or administrative data; and (5) multiple methods (e.g., a combination of the above approaches).

Presented in Table 2.02, the majority of organizations (68%) utilize interviewer observations for the collection of respondent gender, in some form. Of those that rely on observational data, 44% (30% of all firms), require interviewers to determine gender by aural observation only. These organizations *never* ask respondents whether they are speaking to a male or female. A plurality of organizations that use observational respondent gender data (56% (38% of all firms)) implement the ask-assisted method.

Asking respondents to state their gender only when deemed not obvious by the interviewer is the most common way telephone data collection firms collect respondent gender data. Only 15% of firms solely ask respondents to state their gender in telephone surveys. A small amount of organizations (13%) determine gender by multiple methods—using some combination of asking, observing, and checking records. Practices vary from collecting by observation early in the questionnaire in addition to asking for a respondent report at the end, to utilizing administrative records, which ultimately aid an interviewer observation. The four percent of firms that described other methods of collecting respondent gender cited aids such as voter lists, administrative records, and panel membership. Asking for first names to determine gender (still, ultimately, an observation) was another method used. Thus, while nearly seven in ten firms report collecting gender by observation (purely or ask-assisted), there is at least some reliance on observational data for nearly all but the firms that collect only by asking the respondent and those that solely use administrative records.

Table 2.*02*

Data Collection Practices of Respondent Gender

		In telephone surveys, which of the following best describes how your organization most often collects the gender of the respondent?					
		Interviewer observation only	Ask-assisted	By always asking the respondent	By multiple methods	By some other method	Number of cases
Total		30%	38%	15%	13%	4%	289
Sector of Organization	Commercial, non-marketing	43%**	35%	13%	6%**	3%	79
	Commercial, marketing	30%**	33%	17%	15%	5%	138
	Academic, other, not-for-profit	15%**	56%**	13%	17%	0%	54
Age of Organization	15 years old or less	33%	31%	21%*	12%	3%	58
	16 to 30 years old	33%	39%	10%	15%	3%	107
	31 years old or more	26%	44%	16%	11%	4%	103

* *Significant at the 90% confidence level*, ** *Significant at the 95% confidence level*

There are some differences in collection methods across industry sectors, although firms were fairly similar in their practices. This was contrary to an expectation that academic and not-for-profit firms would be more likely to ask respondents their gender and be more rigorous in their data collection approach. Similar proportions across all sectors ask respondents to state their gender. The majority of academic and not-for-profit firms (56%) determine respondent gender using the ask-assisted method; only 13% always ask respondents their gender. However, differences arise when parsing out the observational methods (pure vs. ask-assisted). Commercial, non-marketing organizations are significantly more likely to determine respondent gender by pure observation than their marketing counter-parts or academics. For marketing organizations, 30% collect gender by observation and a similar number (33%) instruct interviewers to ask gender

only when needed. Some firms (17%) ask respondents whether they are a male or female.

I hypothesized that older firms—those that may have carried practices, expectations, and assumptions from traditional face-to-face and mail modes—would be most likely to rely on interviewer observations of respondent gender. However, collecting gender by observation is prevalent among younger firms too. A large group (64%) of data collection organizations that are less than 15 years in age determine respondent gender by some form of observation (solely or by asking when needed) compared to 70% of firms that are more than 30 years old. The use of the ask-assisted method is most popular among older firms. Asking respondents to state their gender is most common among the youngest organizations, though still only one in five do so. Although one might hypothesize that organizations that subscribe to the AAPOR code of ethics would be more rigorous in their methods, only 15% of AAPOR affiliates report always asking respondents their gender compared with 21% of non-affiliates.

Findings discussed above describe the methods firms utilize "most often." Across all telephone data collection firms, as well as within sectors, more than half of the organizations (53%) report that their collection methods vary across projects or clients. The survey did not ask which methods are used secondarily.

A total of 76% of the firms that use observations of gender require interviewers to determine whether they are speaking with a man or a woman with relatively little training. About one-quarter (24%) of organizations stated they provide specific interviewer training on the process. However, among the firms that do train, protocols for most are quite simple. For instance, one firm that denoted collecting respondent

gender by the ask-assisted method included their interviewer instructions. It read: "*CODE WITHOUT ASKING IF POSSIBLE. Are you Male or Female?*" Here, interviewers are encouraged not to ask, unless necessary. Other similar instructions included: "*Listen…ask [gender] as necessary.*" A few firms appear to provide slightly more guidance for their interviewers when observing respondent gender. One noted the following: "*We stress the fact that voices can often be misleading (e.g., respondent with emphysema) and the question should be asked if the gender is at all unclear.*" Another organization wrote: "G*enerally describing how males' pitch is about an octave lower than females', but noting that here is much overlap and if there are any doubts, then the interviewer needs to ask."*

The survey asked firms that use the ask-assisted method what percentage of the time they think their interviewers ask respondents whether they are male or female. The questionnaire did not ask respondents to validate or consult interviewers; thus, these responses are likely a "best guess." Contrary to their widespread belief that asking gender is offensive and usually obvious, these firms estimate that interviewers actually ask gender in 10% of cases (see Table 2.03). Firms who use observational methods were also asked to estimate how accurate interviewer observations of a respondent gender are over the phone; 95.7% was the mean reported accuracy. Upon further dissection of the data, firms were optimistic about the perceived accuracy of their interviewers' gender

observations. The median score was 98% (ranging from 2%-100%)[10] with 13% of the firms believing the accuracy of their interviewers' judgments of respondent gender to be 100%. Another 47% said either 98% or 99%. Only 4% of the observational firms that provided data reported an accuracy score of less than 90%. Firms that collect gender by asking the respondent were also asked about their perceptions of the quality of interviewer observations of respondent gender. They perceived a notably lower accuracy rate of 82.8% (ranging from 50%-99%; median of 90%) in observations of gender. 47% of these firms reported an inaccuracy of less than 90%.

Table 2.03

Frequency Firms Ask Gender and Perceived Accuracy of Observations

	Mean Percent	Number of Cases
What percent of the time do you think interviewers do, in fact, ask the respondent their gender?	10.0%	81
How accurate would you say interviewer observations of a respondent gender are over the phone?		
Firms that collect by observation	95.7%	152
Firms that collect by asking the respondent	82.8%	36

2.2.2. At What Point in a Survey do Firms Collect Respondent Gender?

Among the firms that collect by observation (either purely or ask-assisted), as shown in Table 2.04, 45% most often do so early in the telephone survey as part of the introduction or screening. Although these are subjective measures (as opposed to empirical measures such as length of exposure to respondent's voice in seconds), limited

[10] One firm reported two percent. No other values between two percent and fifty percent were obtained. I recognize that this dramatically reduced the mean score; however, I chose to retain this value for analysis since I could not assume this outlier was not valid.

acoustic information is presumably available to interviewers making these early gender assessments. The majority (53%) of those that use interviewer observations determine respondents' gender at the end of the questionnaire. These observations are likely informed not only by increased exposure to vocal properties but also by hints from spoken language. Only two percent of firms collect respondent gender in the middle of the questionnaire.

Table 2.04

Placement of Interviewer Observation of Respondent Gender

		In your telephone surveys, which of the following best describes where in the survey interviewers most often observe the respondents' gender?			
		In the survey introduction or screening	In the middle of the survey	At the end of the survey	Number of cases
Total		45%	2%	53%	221
Sector of Organization:	Commercial, non-marketing	46%	2%	52%	63
	Commercial, marketing	51%	1%	48%	99
	Academic, other, not-for-profit	37%	4%	59%	46
Age of Organization:	15 years or less	40%	3%	58%	43
	16 to 30 years	43%	2%	55%	88
	31 years or more	54%	1%	45%	78

[*] *Significant at the 90% confidence level,* [**] *Significant at the 95% confidence level*

Here, some industry differences are observed. Denoted by the lack of asterisks, the differences are not statistically significant, however, there could be practical significance to consider. Commercial and marketing firms have higher prevalence than academic organizations of requiring interviewers to determine gender in the introduction or screening. Potentially due to the routine targeting of specific gender markets and quota sampling, half of marketing firms (51%)—compared with 37% of academic or not-

for-profit organizations—instruct interviewers to observe or ask only when necessary whether they are speaking with a male or female respondent early in the survey. Commercial, non-marketing firms follow practices that are similar to those used in marketing sectors; 46% of these organizations require interviewers to observe gender as part of the study introduction or screening while 52% do so at the end of the questionnaire. In terms of placement, the collection of respondent gender does not vary significantly by the age of an organization. Firms that have been established for over 30 years appear to be slightly more likely than younger organizations to collect gender observations early in the survey, however, small sample sizes in these cells likely contribute to the lack of statistical significance. A majority of newer firms favor observing gender at the end of the questionnaire.

Placement of gender observations frequently varies by project or client, as only 25% of organizations consistently collected gender observations in the same location across studies (see Table 2.05). Overall, three quarters of firms (75%) reportedly change at what point interviewers observe respondent gender as required by the study or client. This also varies by industry sector. Fewer academic organizations (66%) than commercial marketing firms (82%) adjust their approach across studies. One-third (34%) of academic and not-for-profits and 26% of commercial, non-marketing firms always ask interviewers to determine respondents' gender in the same questionnaire place.

Table 2.05

Variation of Placement of Interviewer Observation of Respondent Gender

		Does your organization always collect gender in the same place in the questionnaire or does the location vary depending on the project or client?		
		Always in the same place	Placement varies across projects or client	Number of cases
Total		25%	75%	272
Sector of Organization:	Commercial, non-marketing	26%	74%	73
	Commercial, marketing	18%	82%**	131
	Academic, other, not-for-profit	34%**	66%	53
Age of Organization:	15 years or less	33%**	67%	55
	16 to 30 years	18%	82%**	100
	31 years or more	25%	75%	99

* *Significant at the 90% confidence level*, ** *Significant at the 95% confidence level*

2.2.3. Why do Organizations Choose Certain Methods to Collect Respondent Gender?

Organizations were asked to describe the rationale for their chosen methods. Table 2.06 presents the coded responses. Responses that fell into more than one category (e.g., "It is quicker and avoids embarrassment.") were coded into the category first mentioned. Among firms who collect gender by observation (either purely or ask-assisted), the plurality of firms (43%) felt it was offensive and insulting to the respondent. Reactions included "It's too rude to ask," "It would be insulting to the respondent and the interviewer should be able to tell from the voice and/or the responses given to certain questions," "It's a very personal issue and direct questioning can easily alienate a

respondent," and "Do not want to offend/alienate respondent... you create a negative bias on subsequent responses." About one in five organizations (22%) felt it was just simply unnecessary as gender should be obvious. Examples of verbatim responses are: "In most cases, the respondent's sex is obvious (so it's just unnecessary to clarify with the respondent)," and "It's awkward to ask gender because in almost all cases you can identify based on voice." Other firms (14%) noted efficiency savings by way of reducing the length of the questionnaire, cost, and burden for both interviewers and respondents. Respondents stated the following: "To facilitate the ease of administering the questionnaire; to not add unnecessary length to the questionnaire," and "Adding an additional question would be more costly." For seven percent, it is more about how practices have always been done: "Shall we say 'force of habit?'" and "It is a longstanding practice rooted in the assumption that some people might be offended to be asked their gender."

Table 2.06

Rationale for Collecting Gender by Observation

Rationale for collecting gender by observation	Firms that collect gender by observation	Number of cases
Offensive or rude	43%	62
Obvious or unnecessary	22%	32
Cost and improved efficiency	14%	20
Client request	3%	4
Standard practice	7%	10
Other	11%	16

The survey asked firms that obtain gender by directly asking the respondent to provide an open-ended response to explain their rationalizing for their chosen collection method. For 66% the reason was straightforward: accuracy (Table 2.07). Reactions from firms included "*Because interviewer by phone may incorrectly assign gender*,"

"*Eliminate interviewer bias or assumption error*," and *"Gender is not always clear by telephone and in order to ensure consistent measurement, it is always asked. Small time requirement for certainty in data collection."* Other responses cited consistency or a client requirement. Other responses included: *"We conduct IVR research…we have to ask them, because there is no live interviewer who can make the determination."*

Table 2.07

Rationale for Collecting Gender by Asking Respondent

Rational for collecting gender by asking respondent	Firms that collect gender by asking respondent	Number of Cases
Accuracy	32%	11
Certainty	41%	14
Consistency	6%	2
Client request	8%	3
Other	12%	4

2.2.4. How is Respondent Gender Data Used by Organizations?

A series of questions were asked to determine how gender data—those collected by observation as well as those determined by asking respondents—are utilized by firms. Like other forms of paradata, perhaps errors in interviewer observations of respondent gender are inconsequential if their uses—and therefore impact on estimates—are minimal. However, the survey found that gender data serve a variety of functions for many firms. Utilization is most prevalent for three purposes (see Table 2.08): as a substantive variable for reporting and analysis (74% of firms who collect gender via observational methods report they use these paradata for this purpose in all or many studies; 83% of organizations that ask for a respondent report do so); for screening to determine eligibility for participation in the survey (52% of observation firms compared

with 56% of respondent-report firms); and for weighting (46% of observation firms vs. 68% of respondent report firms).

Table 2.08

Uses of Interviewer Observations of Respondents' Gender

	Frequency with which organizations typically use interviewer observations of a respondent's gender for various purposes.		
	For all or many studies	For few or no studies	Number of cases
Firms that collect gender by observation			
As a substantive variable used in analysis and reports based on the survey	72%	28%	193
Screening to determine eligibility for participation in the survey	52%	49%	194
For weighting purposes	45%	55%	194
To assign skip patterns/inform survey logic	25%	75%	194
Other forms of non-response adjustment	19%	81%	194
For interviewer tailoring or accommodation strategies	15%	85%	193
Firms that collect gender by asking respondents			
As a substantive variable used in analysis and reports based on the survey	83%	17%	47
For weighting purposes	68%	32%	47
Screening to determine eligibility for participation in the survey	56%	44%	48
To assign skip patterns/inform survey logic	30%	70%	47
Other forms of non-response adjustment	19%	81%	47
For interviewer tailoring or accommodation strategies	13%	88%	48

One-quarter of organizations (25%) use interviewer observations of gender to assign skip patterns in all or many of their studies, while 30% of those that require interviewers to ask respondents their gender use these data to determine survey skip logic.

Not presented in the table, although investigated, are differences in use by industry sector. Commercial firms, both marketing and non-marketing, rely on interviewer observations of respondent gender to route respondents through survey logic significantly more than academic or not-for-profit firms. Only 15% of academic organizations use these interviewer judgments to assign skip patterns in all or most

studies, compared with 27% of marketing firms. While weighting and other forms of non-response adjustment uses did not differ across industry sector, commercial, marketing firms were more than twice as likely as non-commercial organizations to use gender observations for interviewer tailoring or accommodation strategies. A majority (63%) of marketing organizations report relying on observational data for screening to determine survey eligibility, providing support for the earlier notion that marketing firms use interviewer observations of respondent gender for respondent targeting. Only 27% of academic firms do the same.

There does appear to be a connection between some uses of observational gender data and at what point in the questionnaire interviewers are instructed to observe respondent gender. Among the organizations that collect gender through interviewer observation in the survey introduction or screening, 71% use this information in all or most studies to inform respondent screening and survey eligibility. As a comparison, 36% of firms that ask interviewers to determine gender at the end routinely use the information the same way. Additionally, 37% of firms that require interviewers to observe respondent gender early regularly use this information to inform survey logic and skip patterns. This compares with 13% of firms that observe gender at the end. Additionally, 70% of firms who require their interviewers to collect gender as part of the introduction or screening use these data as a substantive variable in analysis in all of most studies. These findings provide supporting motivation for Chapter 3 of this dissertation which explores how the placement of an interviewer observation predicts gender misclassification and for Chapter 4 which evaluates the impact on survey estimates when

inaccurate gender observations are used in substantive analyses. Forty-nine percent of firms observing early in the survey use the observations for weighting purposes.

2.3. Discussion

This study presented findings from a survey of 289 data collection firms that conduct telephone surveys. The goal was to document common practices in the collection of respondent gender data. Based on the results, while there is significant variation in practice, 38%, the plurality, use a form of observation where interviewers determine gender by aural cues but ask respondents their gender when necessary—ask-assisted. One in three firms collects gender purely by observation; 45% of those doing so early in the questionnaire while 53% observe at the end of the survey. Few organizations (15%) always ask respondents whether they are a male or female. For many firms, gender is observed rather than asked because it is viewed to be offensive to ask or simply unnecessary since it is assumed to be obvious.

Some differences in practices were found across industry sector and age. Commercial organizations, especially those in marketing, were significantly more likely than academic or not-for-profit firms to determine respondent gender purely by interviewer observation. Apart from their collection methods, commercial firms rely on gender paradata to route respondents through survey logic significantly more than academic or not-for-profit firms. In these situations, interviewer misclassification could have notable impacts on final estimates and survey nonresponse. For example, consider a marketing study that has sets of questions designated for men and others for women. If gender misclassification occurs early, some questions that are meant to include only women may have significant measurement error if men are not being properly filtered.

These final survey estimates do not represent women (the targeted population) as men are also included in the data. In this example, survey nonresponse is also a threat if the incorrect survey logic leads to break-offs before collecting the data targeted for men. In other words, a man first receiving a set of questions asking about make-up and other cosmetic products may, in frustration, terminate the survey before receiving his set of questions which ask about facial shaving products.

This study found that many firms who use interviewer observations to obtain respondents' gender require interviewers to do so early in the questionnaire. This potentially exacerbates some measurement problems noted by linguistics researchers. The limited acoustic information and lack of hints from spoken language available in a survey introduction or screening does not allow for ample cues to discern overlapping pitch ranges between males and females. Does allowing more time to hear the respondent's voice improve the quality of their observation? This question is answered in Chapter 3 of this dissertation.

The most commonly cited reason for collecting gender by observation is to protect respondents from being offended. While survey researchers agree that minimizing respondent burden and increasing rapport to avoid breakoffs is an important consideration, the extent to which asking gender is actually offensive is unknown. Telephone survey respondents may understand that this is simply a required question, rather than something of a sensitive or offensive nature. Experimenting with question construction and qualifiers (e.g., *I'm sorry, but I'm required to ask…*) may find a solution to reliably obtain gender data from respondents while relieving the perceived tension. Additional research may find that the threat to measurement error in an observation is far

greater than the actual consequences of asking respondents to state their gender. Along these lines is the need to further explore the issue of trade-off, specifically, the rationale for continuing with observational methods if firms are aware they are not always accurate. Do they believe the offensiveness or extra time required to ask is greater than the impact of incorrect gender assessments?

Finally, future research could address some limitations of this study. First, this questionnaire did not ask firms for a measure of their size (e.g., number of surveys conducted annually, number of employees, etc.); this is perhaps an important variable in understanding industry differences. Second, only firms conducting research in the United States are represented. A cross-national, multi-cultural investigation of this issue is important when identifying industry practices since social norms, values, and expectations vary globally. In some countries, asking a respondent their gender may be expected and even appreciated, while, in others, such questioning may be offensive and culturally unacceptable. Third, the study looked at telephone survey practices in general and did not differentiate between variations within this mode (e.g., approaches in conducting landline vs. cell-phone interviews), causing potential confounding between telephone survey designs and methods. A possible reason why firms change their method depending on the client and project is that different forms of telephone data collection may call for different approaches in the collection of gender. Asking gender in an IVR mode, for example, seems necessary since a live interviewer is typically not part of the administration process.

3. Sources of Inaccuracy in Interviewers' Observations of Respondents' Gender and Its Impact on Nonsampling Errors

This chapter of my dissertation identifies and experimentally tests sources of error in interviewer observations of respondent gender. As discussed in Chapter 2, 68% of survey organizations that perform telephone data collection determine respondent gender by observation—either a pure observation or *ask-assisted*, that is, aided by an instruction for interviewers to ask only if the respondent's gender is not identifiable. Preliminary research (McCulloch et al., 2010) that compared telephone interviewer observations of respondent gender with the presumed true value (respondent report), found an overall gender misclassification rate of 8%. Expanding upon this work and guided by the linguistics literature, here, I conduct an experiment to test two situational predictors of inaccurate judgments of respondent gender in telephone surveys. Respondent and interviewer characteristics were assessed as covariates of misclassification error.

The situation, or environment, in which telephone interviewers determine respondents' gender could vary. Two such situational conditions are explored in this chapter: exposure length (Does how much time survey interviewers have to observe whether they are speaking with a male or a female affect the quality of their judgment?); and the presence of noise (Does the noisy environment of a centralized phone room contribute to observational errors?). Existing literature in linguistics focuses mostly on short vocal segments (e.g., Lass et al., 1976) when evaluating observational errors. Harb and Chen (2005) examined the issue of exposure length. They observed an improvement in overall gender classification accuracy with only small increases in exposure to voice

samples. For one second of listening time, six and a half percent of recorded telephone samples (nine male and nine female voices) were incorrectly identified as male or female. Yet, after five seconds of exposure, misclassification rates dropped to one and a half percent. These rates, however, only represent the outcomes of a single expert coder, which is a notable shortcoming because certain coders could induce greater levels of error or bias than others (U.S. Bureau of the Census, 1972; Baily et al., 1978). Aside from Harb and Chen's work, little is known about how exposure length affects the error in gender observations from the voice alone. In the survey setting, assessing interviewer (coder) variance is important, because many interviewers typically collect data for a study.

For survey practitioners and designers, length of exposure time could be crucial when deciding where to place interviewer observations of respondent gender. Chapter 2 documented that for 45% of firms, gender observations are made in the survey introduction or during the screening process after only limited exposure to the respondents' voice. Few firms determine gender in the middle of the questionnaire while 53% do so at the end, allowing for longer exposure of the interviewer to the respondent's voice and for vocal cues to inform interviewers' judgment. Considering this, it is important to identify what effect, if any, length of exposure to a speaker's voice has on the accuracy of a listener's observation of the speaker's gender.

The second situational condition experimentally tested here is the presence or absence of auxiliary noise. Linguists often use non-noisy, laboratory settings for most of their research. This differs from the situation where many telephone interviewers are making their observations, i.e., a centralized phone room where many interviewers are

talking and creating a noisy environment. However, not all telephone survey organizations conduct their telephone data collection from a centralized phone room. In fact, at least one large firm (Westat) now has the majority of their telephone interviewing staff calling respondents from their home. In this arrangement, the noise level cannot be controlled, although I assume a less noisy environment exists at home than standard phone rooms. This experiment determines whether misclassification is predicted by a noisy/non-noisy environment.

Discussed in detail in Chapter 1, recall that anatomical differences between men and women create distinctive features of gender-specific voices (e.g., males have longer vocal tracts leading to a lower pitch and tend to be less breathy than females) (Graddol and Swann, 1989; Laver and Trudgill, 1979). Socially-driven features of language and speech (e.g., women use more hedges and qualifiers (Lakoff, 1975) can also distinguish male and female voices. However, existing work in linguistics documents various speaker characteristics that contribute to errors in observations of gender. Listeners can fail to make correct gender classifications due to overlaps in pitch, creating a gender ambiguous range (Hess, 1983; Shimamura and Kobayashi, 2001; Ross et al. 1974; Oates and Dacakis, 1983; Graddol and Swann, 1989; Mendoza-Denton and Strand, 1998). Identifying gender from voice is also jeopardized by an individual's ability to adapt, tailor, and manipulate his or her speaking style (Klatt and Klatt, 1990), behaviors (e.g., smoking), or medical conditions (e.g., laryngitis) (Graddol and Swann, 1989; Welham and Maclagan, 2003).

The compromised ability to clearly distinguish male and female voices is supported outside of linguistics in McCulloch et al.'s (2010) research. They found

significant respondent characteristics that were predictors of inaccurate gender judgments. Gender itself was a predictor, finding asymmetric misclassification of women. Race, too, was important as they found greater inaccuracy for African-Americans, and specifically African-American women. Hispanic respondents, although not statistically significant, were also slightly more prone to classification errors.

Interviewers (listeners in this scenario) themselves were also found to be a significant covariate in McCulloch et al.'s (2010) study. Women were more likely to be miscoded by female interviewers than male interviewers. Some support for these findings is evident in the linguistics literature where Nyggard and Queen's (2000) experiment found differential ability of men and women as listeners to identify the gender of a speaker. African-Americans were more likely to be miscoded than non-blacks when being interviewed by non-black interviewers. Experienced interviewers (measured by their total number of completed telephone interviews at MIPO) were more likely to make observational errors than those with less interviewing experience.

I treat respondent gender, race, and interviewer characteristics as covariates when understanding the impact of exposure and noise. For instance, do males and females have a differential impact on interviewers' ability to make correct gender assessments in various noisy environments or under different length of exposure to voice? The McCulloch et al. (2010) findings discussed earlier and in Chapter 1, established base rates of gender misclassification. It is important to note that this chapter is not designed nor intended to provide or compare rates of misclassification to those observed in the preliminary work. Instead, here I look at the effect of experimental conditions (exposure length and noise) on misclassification.

Using the discussed research as a theoretical motivation for the design, a laboratory experiment was implemented to test causal hypotheses, which may explain the error in interviewer observations of respondents' gender in telephone surveys. Only the situational characteristics were tested in the experiment. Respondent and interviewer level measures were treated as covariates in the analyses. Specifically, I address the following questions:

1. Do situational circumstances, exposure length and noise, predict error in interviewer observations of a respondent's gender? In other words, does allowing more time to disentangle gender cues improve observation and does a noisy phone room contribute to errors in observations?
2. Are rater (survey interviewers*)* characteristics, such as gender, race and experience, important covariates when assessing interviewer observations of a respondent's gender in various situational conditions?
3. What respondent characteristics, such as gender, race, and age, are important covariates of error in interviewer observations of a respondent's gender in the situational conditions?
4. What are the interaction effects of the situational, respondent, and interviewer, level predictors on misclassification error?

3.1. Data and Methods

To test the effect of exposure time and noise on quality of survey interviewer's gender judgments, I utilized recordings of previously conducted survey interviews. By experimentally varying the length of exposure I determined whether quality improves with increased listening. Additionally, I experimentally induced the presence or absence

of noise to understand if noisy phone rooms decreased the accuracy of gender observations.

3.1.1. Recordings

This research drew recordings from a pool of 25,567 recordings from past telephone surveys conducted by The Marist Poll between July 2010 and December 2011. In all recordings, the speakers—survey respondents—were adults (18 years of age or older) who resided in the contiguous United States and spoke English. In order to best explain the nature of the recordings and steps of the preparation process, I first detail the creation of the analysis groups. I then describe the preparation of the experimental groups.

Stratification of the recordings to create covariates. As discussed, both preliminary research and the linguistics literature document substantial differences in classification error between male and female speakers and between races. To ensure sufficient sample sizes and the ability to include them as covariates in the models, using data from the studies, I first stratified the recordings along three criteria: gender (males vs. females based on respondent self-report), race (white vs. non-white based on respondent self-report), and initial interviewer gender classification (hard vs. easy).

Stratification by the initial classification allows for the analysis of both 'easy' and 'hard' voices. In this chapter, 'easy' cases means respondents were assigned a correct gender by the original interviewer. In other words, the original Marist Poll interviewer—as part of the survey introduction and screening— accurately observed whether the respondent was a male or female using only aural cues. 'Hard' cases are defined as respondents who were assigned an incorrect gender by the original interviewer. I

recognize that although a case was originally observed with the correct gender, it does not always mean the gender identification was, in fact, easy (e.g., a correct assessment could have been made by a lucky guess). Alternatively, a case that was assigned an incorrect gender observation could have been easily identified as male or female, but an interviewer inputting/coding error caused the incorrect response to be recorded. However, for the purpose of this research, I chose to use this information in the stratification process because it serves as the best available proxy for a difficulty measure and I do not suspect there are large errors in these data that would cause results in this experiment to greatly vary. This 2x2x2 experimental design created eight groups:

(1) Hard white males

(2) Easy white males

(3) Hard white females

(4) Easy white females

(5) Hard non-white males

(6) Easy non-white males

(7) Hard non-white females

(8) Easy non-white females

Within each of the eight groups, thirty-six recordings were chosen at random. Random sampling was conducted by assigning a random number to each case and sorting the file by that random number in ascending order. The first 24 recordings in each group were selected resulting in a total of 192 files included in this analysis. Sixteen recordings needed to be substituted (by taking the next recoding in the randomly sorted file) for two reasons: (1) the recording quality was too poor for analysis or (2) recordings could not be

located. All recordings are unique speakers; speakers only appear in one of the cells in Table 3.01.

Preparation of experimental groups. The next step was to create the experimental groups. To address the goal of testing whether judgment improves with increased listening time, two undergraduate research assistants at Marist College and I prepared the recordings by eliminating all interviewer speech, as the nature of a survey interview requires back and forth dialogue between the interviewer and respondent. This was a manual process—using Praat software—in which we carefully listened to each recording, identified the segments of interviewer voice, and cut them from the original file. Overspeech, where both the interviewer and respondent were speaking at the same time was also eliminated. Second, using the prepared files, the recordings were randomly selected to cut into one of four lengths (again, using Praat): 25% of all recordings were cut to 1 second of speaker voice; 25% of all recordings were cut to 5 seconds of speaker voice; 25% of all recordings were cut to 30 seconds; and the remaining 25% of recordings will present all of the speaker voice.

The rationale for these lengths were motivated by the linguistics literature—where much of the research was conducted using brief exposure (e.g., one second or less)—and findings from the survey of survey firms, which found that organizations make gender judgments both early in the questionnaire as well as at the end. The treatments sought to replicate likely vocal exposure in these situations. For example, a firm that instructs interviewers to observe gender in the survey introduction or screening may ask an interviewer to determine gender in approximately one second. The interviewer may only hear "*hello*" before they are required to determine gender. The

recordings reflect this. The one second segments usually only include one to two words (some examples include '*hello*,' 'yes,' '*I am*,' and '*yeah*'). The five second segments provides raters approximately six to eight spoken words (depending on the rate of speech of the speaker) to make their judgments. Examples from the recordings include: '*yeah, what, I don't vote*,' '*sorry I can't understand you, what was that*,' '*yes, I, what town, yes I am*," '*yeah, he's older, I live in uh*.'

All of the recordings take the voice from the beginning of the original recorded interview. This decision was motivated by the following: (1) survey interviewers are likely to begin their assessments and formation of judgments from the first exposure to the respondent's voice; (2) my assumption that vocal attributes should remain fairly stable throughout the interview; and (3) 45% of firms instruct their interviewers to observe respondent gender as part of the survey screening and introduction process. For seven of the recordings, a different person answered the phone than was selected to participate in the survey. In these cases, the speech segment was taken from the point that the actual survey respondent began speaking.

Data from The Marist Poll also helped inform the exposure lengths for the experiment. Interviewers there observe respondent gender as part of their survey introduction and respondent selection process. To determine how much time Marist Poll interviewers typically have to determine whether they are speaking with a male or female, a time stamp[11] was placed immediately before interviewers observe gender. The elapsed time includes the Marist Poll interviewer's standard introduction and asking two

[11] Voxco interviewing software electronically captured this time stamp data for each case.

questions ((1) *Are you 18 years of age or older?*; (2) *How many adults, aged 18 or older, currently live in your household?)* Using time stamps from 358 interviews, on average Marist Poll interviewers observe respondent gender 52 seconds into the interview (time recorded from the moment someone answers the phone). This time includes the interviewer speech, which is notable because they need to read their standard introduction. It is conceivable that if interviewer voice was omitted there is significantly less than 52 seconds of respondent voice—and perhaps even much less than 30 seconds—for interviewers to make gender judgments. The decision for five seconds of exposure in this experiment means that firms would likely still be in the introduction. The 30 second exposure is likely to exceed the screening and introduction section. The full survey exposure replicates what interviewers hear when making assessments at the end of the questionnaire. Table 3.01 shows the distribution of cases, by race, for each of the exposure length treatment groups.

Table 3.01

Experimental Groups with Race Distribution by Length Assignment of Selected Cases

	Hard white males	Easy white males	Hard white females	Easy white females	Hard non-white males	Easy non-white males	Hard non-white females	Easy non-white females
1 second exposure	6 white	6 white	6 white	6 white	1 Black 3 Hispanic 2 Asian	4 Black 1 Hispanic 1 Asian	4 Black 2 Asian	3 Black 2 Hispanic 1 Asian
5 second exposure	6 white	6 white	6 white	6 white	4 Black 1 Hispanic 1 Asian	2 Black 2 Hispanic 2 Asian	5 Black 1 Asian	2 Black 1 Hispanic 3 Asian
30 second exposure	6 white	6 white	6 white	6 white	3 Black 3 Hispanic	3 Black 1 Hispanic 2 Asian	4 Black 2 Hispanic	3 Black 3 Hispanic
Full survey exposure	6 white	6 white	6 white	6 white	6 Black	4 Black 2 Hispanic	4 Black 2 Asian	4 Black 1 Hispanic 1 Asian

Noise, the second situational experimental condition, was then randomly assigned to three of the six recordings in each of the cells in Table 3.01. I chose to measure noise

dichotomously (presence or absence of auxiliary noise) as opposed to levels of noise (e.g., high, medium, low) because there is not yet existing research to provide a framework or measurement of various levels of phone room noise. Plus, practical applications of this work would be better suited to address centralized versus decentralized data collection facilities (as opposed to somewhat noisy versus very noisy phone rooms). There is no evidence or the ability to ensure that a de-centralized arrangement always means the auxiliary noise inherent in many phone rooms is eliminated. However, for the purpose of this work, I assume that de-centralized data collection situations would be absent of the typical noise in a busy, centralized facility.

To implement this variable, an audio recording of the 'noise' (known as a 'mask' in linguists) during one of The Marist Poll's active interviewing sessions was taken. The room contains 48 interviewing stations, all occupied at the time of the recordings. Additionally, approximately 12 other people were in the room performing various supervising and monitoring functions. A recording device was placed high in the center of the room and captured the background noise. To simulate the noisy signal in some phone rooms, the 'noise' was overlaid on half of the recordings (97 randomly selected cases) within each cell (thus, three out of the six recordings in each cell contained the noise). The purpose of randomly selecting cases to have the noise overlay within each of the experimental groups was to ensure that the covariates did not differ by chance (e.g. prevented all of the cases with the noise from being African-American females). This process was done individually for each recording using Praat.

3.1.2. Participating Raters

Twenty-seven undergraduate students who work at The Marist Poll at Marist College in Poughkeepsie, New York served as the study participants, the raters. All were trained telephone interviewers with at least one semester of interviewing experience. Similar to the Lass, et al. studies (1976; 1978; 1979) 14 raters were female and 13 were male; within the females, half were white and half were non-white. Among the male raters, seven were white and six non-white. This design decision enabled me to determine whether certain raters, of specific gender and race, were more adept at making gender judgments. To allow for the analysis of another interviewer-level covariate, I knew the number of shifts a rater had ever worked for The Marist Poll as a proxy for experience. Total shifts ranged from one[12] to 148 among the participating raters; the mean was 28; the median was 17. None of the participants had any previous or additional interviewer experience outside of The Marist Poll.

Much of the research conducted in linguistics is in an experimental lab setting in which the observation and classification of the speaker's gender is the central task. For survey interviewers, this is one of many tasks they need to complete quickly in order to move to the next question. Thus, the attention and thought applied to the task in a survey setting may be considerably different and affect error rates. To prevent raters from unnaturally focusing on their determination of the speaker gender, raters were told that the study was being conducted to evaluate how accurately interviewers are able to

[12] This rater worked only one shift during their first semester of interviewing.

determine attitudinal characteristics about respondents, such as voting behavior and political party identification and that this information could help researchers supplement data for survey nonrespondents. Raters were not made aware of the actual research goals. They were told that, as part of their normal process, they would need to observe the respondent's gender.

3.1.3. Implementation

The study was conducted in The Marist Poll's centralized telephone interviewing facility where participants sat at interviewing stations. Using a within-rater design, each rater listened to and coded a subset of 64 cases: 2 from each of the 32 cells shown in Table 3.01. Nine raters evaluated each recording to evaluate intra-coder reliability.

The 64 randomly selected recordings were placed on each of the raters computers. Wearing dual earpiece, non-noise canceling headsets, participants listened to each recording and completed a questionnaire[13] immediately after each case. The questionnaire included the following items:

1. Do you guess the speaker is:
    a. Definitely a male
    b. Probably a male
    c. Definitely a female
    d. Probably a female
2. Do you guess the speaker is:
    a. Definitely White
    b. Probably White
    c. Definitely Black or African-American
    d. Probably Black or African-American

[13] Qualtrics software was used to collect the data. All questions appeared on a single page and accepted only one response.

    e. Definitely Latino or Hispanic
    f. Probably Latino or Hispanic
    g. Definitely Asian
    h. Probably Asian
    i. Some other race (please specify)
3. In which of the following age categories would you say the speaker is?
    a. Between 18-29
    b. Between 30-44
    c. Between 45-59
    d. Between 60-74
    e. Between 75-89
    f. Over 90
4. Do you guess the speaker is:
    a. Definitely registered to vote
    b. Probably registered to vote
    c. Definitely not registered to vote
    d. Probably not registered to vote
5. Do you guess the speaker is:
    a. A Republican
    b. A Democrat
    c. An Independent

Raters were only permitted to listen to the recording once; thus, no rewinding or replaying of the voice segments was allowed. The rationale for this decision was to replicate a typical situation in which interviewers make observations. None of the original interviewers participated in the experiment, leaving no chance that the same individual evaluated a recording twice. With each of the 192 recordings evaluated by 9 raters, this process resulted in a total of 1,728 cases available for analysis.

3.2. Analysis Methods and Hypotheses

Several methods were used to present the results of the experiment. Stata 12 was used for all analyses. Hypotheses, accepted or rejected by the discussed analysis techniques, are discussed in this section.

3.2.1. Analyses

Descriptive statistics (univarite and bivariate tables) were first used to illustrate the amount of error in the data overall, for each of the experimental treatments and stratification groups. Recall that raters documented speaker gender on a four point scale (definitely a male, probably a male, probably a female, definitely a female). The gender misclassification error discussed in this chapter was calculated by combining the inaccurate definitely and probably values. For example, 45.37% classification error for hard non-white males is the sum of 35.65% (definitely a female) and 9.72% (probably a female). Accuracy was based on the respondent report (assumed to be the true value) from the original interview.

Using logistic regression, I extended the bivariate summaries to better understand the predictors of inaccurate gender classification.[14] For all models, rater error of gender observations serves as the dependent variables (where 0 = no error and 1 = error). The experimental conditions (exposure length and presence of noise) serve as explanatory variables for the situational models. Other covariates in the model included difficulty of the interview (easy, hard) and respondent level characteristics (gender). Other models showing the effects of rater and respondent level covariates on gender classification error omitted the experimental variables. The logit model for exposure length can be represented as:

[14] Logistic models were fit using the *xtlogit* command in Stata 12.

$$\Pr(guesserror = 1) = F(B_0 + B_1 length1\,\text{sec} + B_2 length5\,\text{sec} + B_3 length30\,\text{sec} + B_4 hard + B_5 male)$$

I chose to fit a linear probability model (LPM) to ease interpretation of the coefficients (Mood, 2010) when including interaction terms. Error between the rater observation of the speaker gender and respondent report obtained in the original interview (where 0 = no error and 1 = error) was still the dependent variable in these models. Respondent demographics and interviewer covariates were also added as predictors. The clustering of interviewers was addressed by the inclusion of random interviewer effects.

Covariates. Instead of the original correct/incorrect gender assignment as a proxy for difficulty, another stratification variable could have been the respondents' actual vocal pitch (measured in hertz). From linguistics, we suspect that much of the gender misclassification comes from confusing pitch levels (e.g., men with high pitch are assumed to be female). However, for the purpose of this work, I chose not to stratify in this way for two reasons: (1) pitch values are not typically or routinely available to survey practitioners to make determinations of case difficulty and (2) we know little about how phone rooms affect pitch measurement. Nevertheless, I included this measure as a covariate to help explain misclassifications of respondent gender.

Median pitch values (rather than means, as they are less sensitive to errors in the pitch estimation algorithms) were assigned to each recording by the Praat software. Praat's default assessment of pitch ranges is from 75 to 500Hz. The means that Praat assesses any pitch signals within this range, not differentiating or tailoring its approach for male and female signals. However, the precision of the Praat output for the

assessment of Hz values is increased when adjusting for known gender specific values.[15] These known values came from the self-reports of gender in the original interview. The minimum pitch value that could be assigned by the software was set to 75 Hz; the maximum was 300 Hz. For the female speakers, the values were manually adjusted to a minimum of 150 Hz and a maximum of 500 Hz. Conflicting median pitch values were documented for nine cases in which I determined the most appropriate value to retain.[16] Across all male recordings, the mean pitch was 151.3 Hz; for females the average vocal pitch was 210.7 Hz. Table 3.02 presents the average pitch for each of the experimental groups. Generally, the male speakers have lower pitch than females. However, when comparing the hard and easy columns, some patterns emerge, especially among males. Looking at the average for all files (regardless of length), easy white males have a notably lower median pitch than hard white males. The same is true for nonwhite males. There are not large differences within the female groups. Easy white females have nearly the same median pitch as hard females. Non-white female values are also similar.

[15] Phenomena such as finding pitch when there is not any, not identifying pitch when there is, pitch doubling, and octave jumping induces errors in Praat's assessment of pitch measurement. Common practice is to address this by manually changing the default measurement range.

[16] Conflicting values occurred due to coder inputting errors (e.g., copying and pasting the wrong value from the Praat output screen into the spreadsheet).

Table 3.02

Median Pitch for Each Experimental Group

	Hard white males	Easy white males	Hard white females	Easy white females	Hard non-white males	Easy non-white males	Hard non-white females	Easy non-white females
All files	180.16	134.89	211.70	211.22	155.42	134.82	209.22	210.68
1 second exposure	186.85	137.27	183.31	218.58	169.13	132.28	227.79	215.67
5 second exposure	179.09	142.40	229.05	219.02	182.43	136.20	207.80	213.30
30 second exposure	179.09	127.14	204.75	205.60	138.59	144.55	195.18	214.45
Full survey exposure	181.69	132.77	229.67	201.67	131.53	126.25	206.12	199.31

Rather than treating pitch as a continuous variable[17], I recoded the values to create three categories for the males and females: below normal, normal, and above normal. Boone (1997) suggests that males have an average pitch of 120 Hz (typical range of 80–150); females average 220 Hz (typical range of 175–250). Using these as a guide and considering the distribution of vocal pitch among the selected recordings, I categorized males' pitch into the following: (1) Below normal: 90 Hz–120 Hz; (2) Normal: 121 Hz–140 Hz; and (3) Above normal: 141 Hz–243 Hz. For females, I created the following categories: (1) Below normal: 159 Hz–210 Hz; (2) Normal: 211 Hz–230 Hz; (3) Above normal: 231 Hz–424 Hz. I recognize that the chosen categorization is slightly skewed, according to the average ranges suggested by Boone (e.g., the 120 Hz average for males

[17] Analysis was initially conducted treating pitch as a continuous variable but it appeared that there was not a linear relationship (more of a curvilinear) between pitch and gender misclassification.

is technically part of the below average category). Due to the oversampling of difficult cases, there were many more recordings with higher than normal pitch ranges for males and few below average. Similarly, for females, there were more recordings with lower than average pitch. Given the distribution, in order to allow for this analysis (maintaining as equal sample sizes in each group as possible), adaptations were made. I do not, however, expect this design decision to impact findings related to pitch extremes. The above normal category of male and the below normal of females pitch should still show higher gender misclassification. All models using pitch as a covariate include this categorical variable.

Rater/interviewer-level covariates included in the analysis are: (1) gender (male, female); (2) race (White, Black, Hispanic, Asian); and (3) experience (measured continuously by the number of interviewing shifts). Gender and race were obtained from rater self-reports; experience was obtained from Marist Poll administrative records.

Speaker/respondent-level covariates included in the analysis are: (1) gender (male, female) and (2) race (White, Black, Hispanic, Asian). All respondent-level independent variables were obtained using self-reports in the original interview, assumed to be true-values in this paper.

3.2.2. Hypotheses

Given findings from preliminary work and research in linguistics, I expected to find the following results: (1) gender classification by observation would improve with increased exposure to the voice, if more time to disentangle confusing pitch signals was allowed; (2) the presence of noise would create a more distracting situation for raters, increasing misclassification; and (3) hard cases—cases that were originally assigned an

incorrect gender by the interviewer—will continue to be miscoded because confusing gender signals are not easily distinguishable by another rater.

I anticipated finding the following interaction effects: (1) greater misclassification would be found among recordings of shorter length that include the noise because limited exposure would be more vulnerable to noisy signals (that is, I expected to find an interaction effect of exposure length and noise); (2) exposure length by gender (although I was unsure of the direction, I could image that males and females might be identifiable at different lengths of exposure); (3) exposure by the difficulty of the case, because I expected to find that difficult cases would take longer to correctly classify; and (4) there would be less error when the gender and/or race of the interviewer matches that of the respondent.

3.3. Results

Remembering that in the study design, 50% of the recordings were originally misclassified by survey interviewers, creating an oversampling of difficult voices. This means that the error rates observed here are not comparable to a respondent pool obtained in typical RDD telephone surveys, nor the preliminary research. Thus, rates of misclassifications will be different in this study than those presented in Chapter 1. In this experiment, across the entire sample 20.26% of all recordings were assigned an incorrect gender by the interviewers. Only 2.09% of the easy cases (those that were originally correctly coded) were misclassified in the experiment. Nearly four in ten (38.40%) of the hard cases (those that were originally miscoded) were again miscoded by a different interviewer.

The recordings were clustered by the rater. Estimating an empty model (which contains no covariates) to examine the effect of the clustering of error within interviewers, I found that gender misclassification error did not cluster around certain raters (the intra cluster correlation was zero and the random effects was not significant). In other words, specific raters were not driving the misclassification. As shown in Figure 3.01, error was distributed across all raters. This is contrary to my hypothesis, motivated by McCulloch et al. (2010)), which found that some interviewers were more likely to make errors in their gender assessments than others. It is, however, consistent with other work finding higher rater agreement (e.g., Oksenberg et al. (1986)).

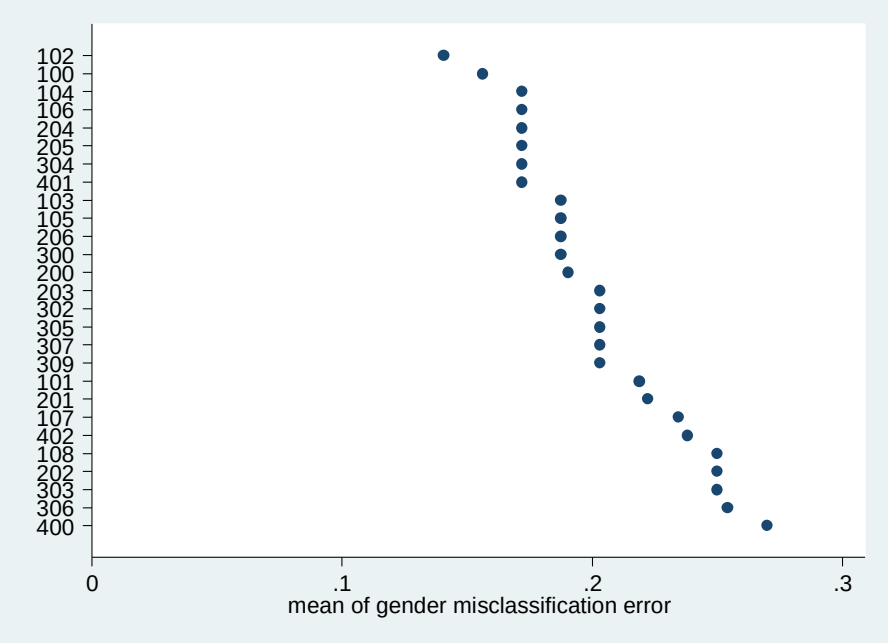


Figure 3.01. Percentage of Error Across Raters; n=64 per rater; Indicates that, for example, Rater #400 misclassified approximately 27% of all recordings.

3.3.1. Situational Predictors

To isolate the effects of the two experimental conditions—length of exposure to the respondent's voice and the presence of noise—the results are discussed below

separately. Significant respondent-level and interviewer-level covariates are then addressed. Interaction effects are presented last.

Length of exposure. As hypothesized, exposure length improved the accuracy of gender observation. As shown in Table 3.03, 28.50% of all 1 second recordings were miscoded. A steady decline in misclassification is observed as time increases: 22.51% of recordings that were 5 seconds in length were assigned an incorrect gender; 17.82% of those with 30 seconds of respondent voice were misclassified. Yet, gender misclassification did not completely disappear with full exposure to the speaker's voice. Even after hearing the entire survey recording, 12.27% of cases were still incorrectly observed.

Table 3.03

Percentage of Gender Classification Error by Exposure Time

	1 second	5 second	30 second	Full exposure
Overall proportion of misclassified cases	28.50	22.51	17.82	12.27

Figure 3.02 shows the mean interviewer gender classification error across the four exposure lengths by difficulty of the case. It is clear that easy cases, regardless of the length of exposure, induced little error when another interviewer observed the voice. Increased listening time had little effect on interviewer's ability to correctly observe the respondent's gender because misclassification was minimal even at one second of exposure. Hard cases continued to be difficult to discern, although length of exposure made a big difference in misclassification rates. The more time an interviewer was exposed to the respondent's voice, the less error was observed in their gender

observations. For instance, 56% of hard cases with 1 second of exposure were assigned an incorrect gender, compared with 31% of those after 30 seconds of listening time. Although significantly less than shorter exposure lengths, the error remained notable (24% misclassification) even after hearing the all respondent voice in the interview. Overall, this visualization showed support for one of my interaction hypotheses (exposure by difficulty of the case).

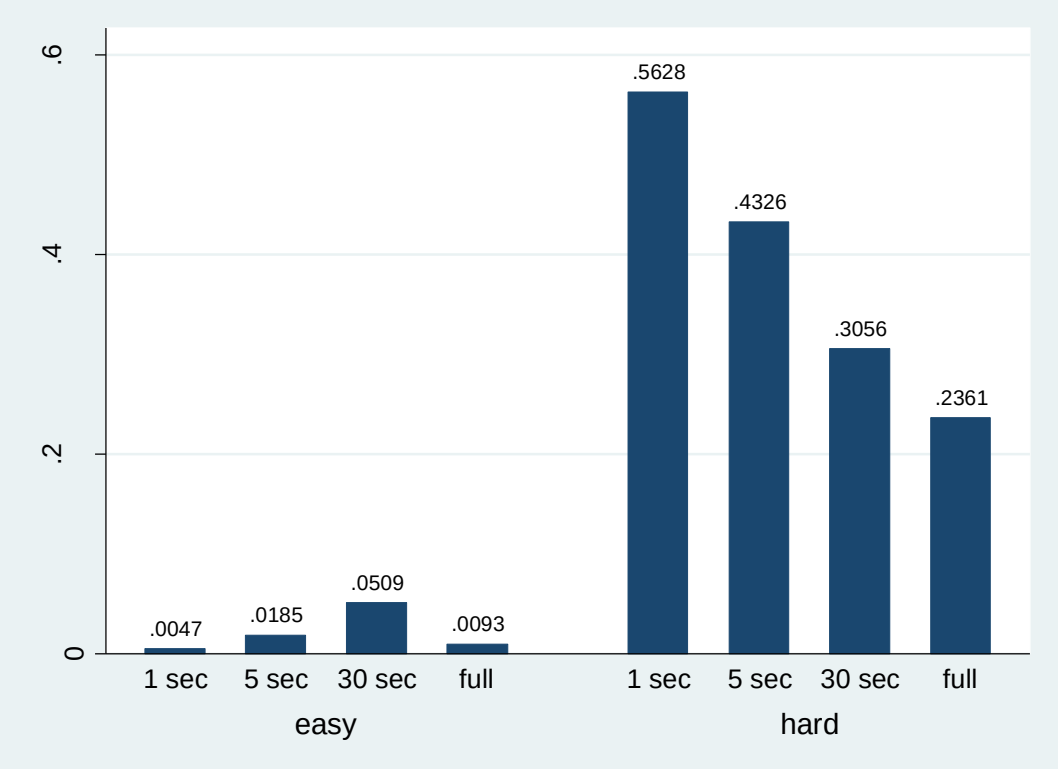


Figure 3.02: Mean Observational Error by Length of Exposure for Easy vs. Hard Cases.

A logistic regression model was used to predict the probability of being misclassified given the four experimental lengths of exposure. The speaker's true-value of gender and the difficulty of the recording were used as covariates. Table 3.04 presents odd ratios. I found that after controlling for gender and difficulty, gender was 3.75 times more likely to be misclassified with only 1 second of exposure to the speaker's voice

when compared with full exposure and recordings of 30 seconds in length were 1.68 times more likely to have an incorrect gender observation. Significant covariates emerged. Recalling that difficult voices were oversampled, these selected males were 1.5 times more likely to be misclassified. The difficulty of the case had a sizable influence on gender misclassification, with hard recordings being 32 times more likely to be miscoded.

Table 3.04

Random Effects Logit Model Results of Exposure on Observational Error

	Odds Ratios (n=1723)
Length of Exposure (Reference category: Full exposure)	
1 second	3.750***
	(.765)
5 second	2.443***
	(.502)
30 seconds	1.681**
	(.352)
Gender (Reference category: Females)	1.538**
	(.213)
Difficulty (Reference category: Easy)	32.098***
	(8.051)
Constant	.008***
	(.002)
Observations	1,723
Sigma_u	.0007
Sigma_e	0.398
Rho	1.43e-07

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

Using the coefficients (rather than odds ratios) in the above logistic regression model, margins were calculated that contrasted the predicted values of gender misclassification for each exposure length for men with women, by the difficulty of the

case, on the probability scale.[18] Margins are another way to show how gender misclassification changes from 0 to 1, after controlling for exposure length, gender, and difficulty. I plotted the values to create a visual depiction of the effects (Figure 3.03).[19] Looking at the plot, there appear to be clear main effects when dealing with this set of selected recordings: (1) gender misclassification decreases as exposure length increases, especially after 30 seconds of exposure (the slope of the line changes slightly at 30 seconds); and (2) easy cases have much lower misclassification than hard cases. The confidence intervals among the hard cases do overlap for males and females, indicating no statistically significant differences within strata. Here, because no interactions were included in the model, the lines do not intersect. Later models presented in this paper address possible interaction effects.

[18] Coefficients obtained using *logit* command with a *vce* option to account for the clustering of interviewers. Obtained using the *margins* command in STATA12

[19] Produced using the *marginsplot* command in STATA12

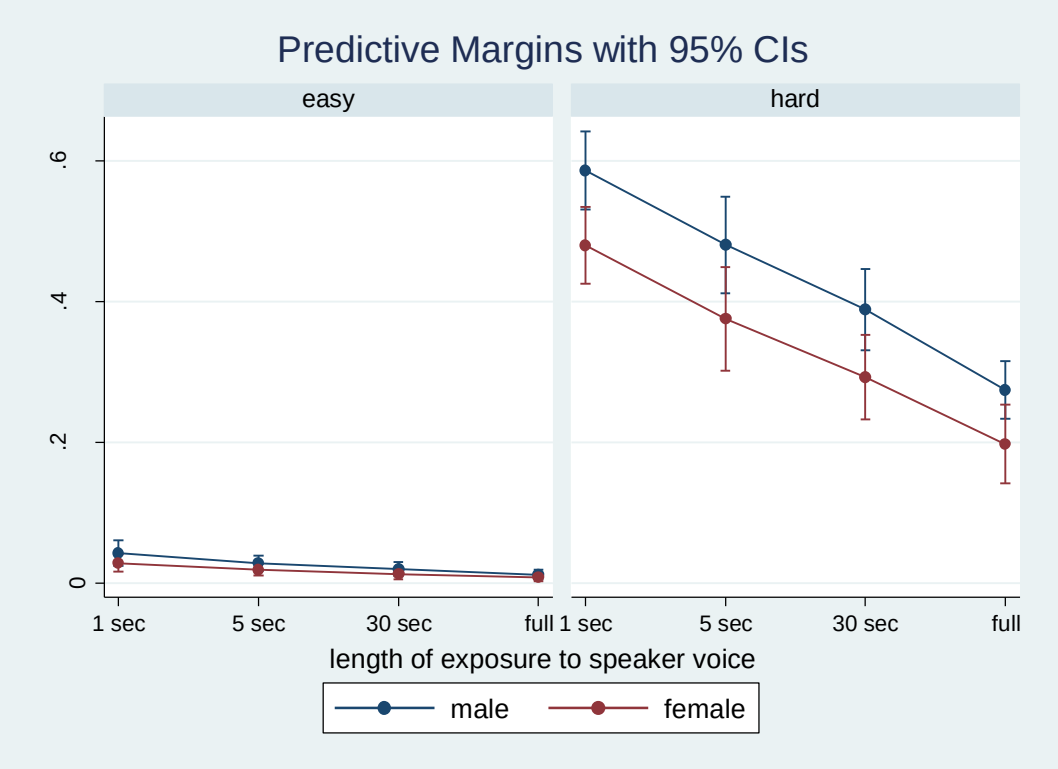


Figure 3.03: Predicted Probability of Misclassification Error by Gender Across Difficulty of the Case

Presence of noise. Overall, as shown in Table 3.05, 17.61% of all cases with the noise overlay were miscoded; 22.91% of the cases without noise were assigned an incorrect gender. This is contrary to my expectation that noisy signals would be prone to greater gender misclassification. One possible explanation for this finding is that when faced with a more difficult signal, interviewers concentrate harder, thus improving the quality of their observations. Other possible explanations are included in the discussion section of this paper.

Table 3.05

Percentage of Gender Misclassification Error by Presence of Noise

	Noise	No noise
Overall proportion of misclassified cases	17.61	22.91

Table 3.06 displays the findings of another logistic regression model using interviewer gender classification error as the dependent variable and presence of noise as explanatory variable. Gender and difficulty of the case are included as covariates and the clustering of interviewers are accounted for in the model. Suggested in the bivariate analysis, the presented odds ratios confirm that the presence of noise actually improves the quality of gender observations. Raters were 1.5 times more likely to correctly observe the respondent's gender when presented with recordings that included the noise overlay.

Table 3.06

Random Effects Logit Model Results of Noise on Observational Error

	Odds Ratios (n=1723)
Presence of Noise (Reference category: No Noise)	1.512 **
	(.213)
Gender (Reference category: Females)	1.512**
	(.205)
Difficulty (Reference category: Easy)	30.059***
	(7.486)
Constant	.013***
	(.003)
Observations	1,723
Sigma_u	.0007
Sigma_e	0.398
Rho	1.78e-07

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

3.3.2. Rater Characteristics

Table 3.07 shows that none of the rater-level covariates were significant predictors of gender misclassification, as indicated by the lack of asterisks. Despite other work

suggesting otherwise, I found no evidence that male and female survey interviewers were disproportionately likely to wrongly observe respondents' gender. Race of the rater also showed no meaningful differences. Experience—measured by the rater's total number of shifts worked—is not a significant predictor in this experiment. Recalling from McCulloch et al.'s (2010) work, experienced interviewers were actually more likely to misclassify respondent's gender. One possible explanation for the lack of significant experience findings here was the experimental setting. That is, raters, regardless of their interviewing experience were participating in a new task, a task that all participants concentrated on equally.

Table 3.07

Random Effects Logit Model Results of Interviewer Characteristics on Observational Error

	Odds Ratios (n=1723)
Interviewer Gender (Reference category: Females)	1.201
	(.148)
Interviewer Race (Reference category: White)	
Blac	1.170
	(.163)
Hispanic	1.233
	(.242)
Asian	1.153
	(.269)
Interviewer Experience	.999
	(.000)
Constant	.217***
	(.028)
Observations	1,723
Sigma_u	.0003
Rho	3.26e-08

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

3.3.3. Respondent Characteristics

Recall that the stratification of recordings by gender and race allowed for the analysis of respondent-level covariates such as gender and race. Overall, regardless of the difficulty of the case, 22.91% of the male recordings were wrongly observed female by the interviewers; 17.61% of the female cases were miscoded. When looking at the original interviewer's observation, 44.55% of hard men were again wrongly observed a female in this experiment, while 32.25% of hard females were wrongly classified. Few easy cases were incorrectly observed (1.17% of males; 3.01% of females). The equal allocation of male and female incorrect and correctly coded recordings means this study is not designed to provide estimates about the differential measurement error between males and females in a naturally occurring setting. The design, however, allowed for the investigation of who was likely to be misclassified again (systematic misclassification) given certain conditions.

Table 3.08 presents the gender assigned by the raters for the stratification groups. The assignments into definitely and probably categories provides some evidence that using the 'ask-assisted' measurement approach discussed in Chapter 2 (instructing interviewers to confirm the respondent's gender when unsure) would not have improved the misclassification error. Presumably the only cases where gender would have been confirmed by the interviewer in the ask-assisted method were those assigned with a *probably*.

Table 3.08

Percentage of Gender Classification by Experimental Group

	Hard white males	Easy white males	Hard white females	Easy white females	Hard non-white males	Easy non-white males	Hard non-white females	Easy non-white females
Definitely a male	40.47	92.09	26.51	.93	42.59	91.59	18.06	3.24
Probably a male	15.81	6.98	6.05	.46	12.04	7.01	13.89	1.39
Definitely a female	34.88	.47	52.56	95.37	35.65	0	42.59	91.20
Probably a female	8.84	.47	14.88	3.24	9.72	1.40	25.46	4.17

In order to incorporate the measure of vocal pitch (Hz) into the model, I fit logistic models separately for males and females, because I assumed the effects would be opposite by gender (e.g., misclassification would be greater when pitch was higher for males; misclassification would be higher when pitch was lower for females). I also chose to model males and females separately to identify whether overall respondent-level effects discussed above were more prominent in one sex verse another. The difficulty of the case was used as a control. Table 3.09 shows the odds of being misclassified, where Model 1 includes male recordings and Model 2 includes female recordings. Vocal pitch was a clear predictor of gender classification error among male respondents. Those with above normal voices were 7.4 times more likely to be incorrectly observed as a female compared with males who had below normal pitch. For women, however, the findings were contrary to my expectation. Women with higher than normal pitch were slightly more likely to be misclassified a male than those with lower pitch. Also notable in this model are the race findings. While the misclassification of females was not predicted by

the race of the speaker, for males, race was a significant characteristic. Consistent with preliminary work, African-American respondents were 4.8 times more likely to have an incorrect gender observation than white respondents. Asian speakers were also more likely to be miscoded.

Table 3.09

Random Effects Logit Model Results of Respondent Race and Pitch on Gender Observational Error by Gender of the Respondent

	Odd ratios: Model 1: Males (n=860)	Odds ratios: Model 2: Females (n=863)
Respondent Race (Reference category: White)		
African-American	4.824***	1.326
	(1.441)	(.284)
Latino	1.204	.504
	(.451)	(.280)
Asian	7.089***	.971
	(3.813)	(.327)
Vocal Pitch in Hertz (Reference category: Below normal)		
Normal	.068**	1.207
	(.057)	(.329)
Above normal	7.405***	1.928*
	(3.092)	(.472)
Difficulty (Reference category: Easy)	76.77***	14.579***
	(40.604)	(4.459)
Constant	.001***	.026***
	(.001)	(.008)
Observations	860	863
Sigma_u	.0004	.0004
Rho	.003	4.19e-06

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

3.3.4. Interaction Effects

As mentioned in my hypotheses, I did expect to find a number of interaction effects between various situational, interviewer, and respondent level characteristics. Through the investigation of the above discussed main effects, some anticipated interactions became unnecessary to test. For instance, I hypothesized an interaction between exposure length and noise (greater misclassification of noisy, short recordings),

however, given that I found that noise actually improved gender classification, I omitted these terms from any models.

Table 3.10 presents the findings when all relevant situational, respondent, and interviewer level factors are fitted (both as main effects and interactions) in an LPM model. Denoted by asterisks, I do, again, observe some significant indicators of probability of making a gender observation error, especially by the experimental variables – exposure length and noise. Looking at the levels of exposure, the 1 second coefficient indicates that when presented with a recording of 1 second in length, it increases the probability of making an incorrect gender assessment by 22.8% compared with full exposure to the respondents' voice. Gender observational error decreases with more exposure. The noise coefficient means that when presented with a noisy signal, the probability of gender misclassification decreases by 5.6%. After controlling for other variables, this model now shows that the main effect for the differential misclassification of males is no longer significant. African-American speakers are slightly more likely to be misclassified than whites, whereas Hispanics have a greater probability of having a correct gender observation than white speakers.

Recall in Figure 3.03 that the slopes of the lines for males and females changed slightly from 30 seconds to the full exposure lengths. Given this, and my expectation that there would be some gender differences in misclassification by length of exposure, this interaction was included in the model. Significant coefficients were found.

Table 3.10
Random Effects Model (LPM)

	(1) Model 1 b/se
Length of Exposure (Reference category: Full exposure)	
1 seconds	0.228***
	(0.03)
5 seconds	0.123***
	(0.03)
30 seconds	0.141***
	(0.03)
Presence of Noise: Noise (Reference category: No noise)	-0.056***
	(0.02)
Respondent Gender: Male (Reference category: Female)	0.051
	(0.04)
Respondent Race (Reference category: White)	
African-American	0.049*
	(0.02)
Hispanic	-0.093***
	(0.03)
Asian	0.024
	(0.03)
Interaction: 1 Second Exposure & Male (Reference category: Full exposure & Female)	-0.112*
	(0.05)
Interaction: 5 Seconds Exposure & Male	-0.030
	(0.05)
Interaction: 30 Seconds Exposure & Male	-0.136**
	(0.05)
Difficulty of Case: Hard (Reference category: Easy)	0.278***
	(0.02)
Interaction: Hard & Male (Reference category: Easy & Female)	0.158***
	(0.03)
Rater Gender: Male (Reference category: Female)	0.032
	(0.02)
Interaction: Respondent Male & Rater Male	-0.005
	(0.03)
Constant	-0.082**
	(0.03)
sigma_u	0.000
sigma_e	0.349
Rho	0.000

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

3.4. Discussion

Through the implementation of a randomized experiment, this chapter found that gender misclassification is significantly reduced when interviewers have longer exposure to respondents' voices. However, the difficulty of the case is a strong interacting

variable; hard cases still show high classification error even after raters are exposed to all of the respondent's voice. Additionally, I showed that the noisy environment of centralized phone rooms may not interfere with interviewers' ability to make correct observations of respondents' gender. The experiment found that, although rater characteristics such as gender, race, and experience did not predict gender misclassification, respondent attributes—such as gender and race—did significantly impact the accuracy of a rater's gender observations.

As mentioned early in the text, misclassification rates observed in this chapter are not comparable to the rates of gender misclassification established in Chapter 1. The goal here was not to give misclassification rates but, instead, to identify the characteristics that explain misreporting with the experimental variables (exposure length and noise) – not the stratification variables (e.g., gender). Future work could incorporate this by performing a weighted analysis to control for the probabilities of selection.

This research does have limitations, which should be addressed in future research. First, the raters were quite homogenous, as Marist Poll's interviewer pool was comprised of undergraduate students enrolled at Marist College. While the linguistic literature does not note correlations between the age of the speaker and that of the listener with respect to the quality of their judgments, future research should consider including raters of various ages. Secondly, the measurement of gender on a four-point scale was something that was different for the participating raters, possibly drawing more attention and atypical focus to this task. Third, the recordings only included the respondent voice, which means cues from spoken language or conversational interactions with the interviewers could not aid in the rater's determination of gender. We know survey

interviewers have more than voice available to make observations about the respondents. However, I felt it was important to first isolate only one cue. Future research might include both respondent and interviewer voices to better understand how components of conversation, as well as voice alone, affect the accuracy of gender observational errors. Finally, each recording appeared only in one of the cells. The rationale for this design decision was motivated by the desire to include a wider range of voices (thus, allowing for the analysis of greater variability in pitch). However, future research may want to use the same recording in the different length conditions (e.g., thus, there would be four version of each recording – 1 second, 5 seconds, 30 seconds, and full survey length), to fully control for other vocal characteristics.

The finding that the presence of noise actually improves gender observations raises many questions. Given that half of the recordings were randomly selected within each analysis group to receive the noise overlay, it is unlikely that the greater misclassification of non-noisy cases was due to a clustering of observed characteristics (e.g., race and gender). However, although we do not have data to investigate this issue, it could be that there is some clustering of unobserved characteristics (e.g., other vocal attributes such as jitter) which affected findings.

There are three possible explanations of what may be driving the misclassification of non-noisy cases associated with interviewers. First, it is common to see interviewers in centralized phone rooms leaning into their stations and pressing the headsets tight to their ears when having trouble hearing. Perhaps, when challenged by a noisy signal, they give greater concentration, thus improving the quality of their observations. A second explanation is that interviews become simply accustomed to their noisy work

environment. Does the chatter of other interviewers become like "white noise" eventually? If so, is it more distracting or disturbing when the room is silent, prompting interviewers to fail to attend to the details of the interview and focus instead on the lack of auxiliary sound? A third explanation is whether the noise causes interviewers to make judgments only from the extremes of the respondents' voices. The noise may blur or mask the voice coming from the middle ranges, thus, interviewers hear only the signals that stand out (those that may inform a higher accuracy of gender assessments). For survey practitioners, experimenting with various conditions of noise levels seems like an important area when evaluating sources of data collection errors.

Although the study does find notable improvements in observations as exposure increases, the error never fully disappears. In fact, contrary to my expectations, the error is still notable even after full exposure to the interviewer. Recall that Chapter 2 found that approximately half of survey firms who collect respondent gender by observation do so at the end of the questionnaire; 45% require interviewers to observe the respondents' gender as part of the survey introduction or screening. The findings in this experiment certainly raise concerns for quality of the observations made early in the interview; however, they challenge any researchers assumptions that extended exposure eliminates substantial errors. While this may be true for some easy voices, others remain challenging for interviewers to differentiate gender despite the length of exposure. The implications for researchers and practitioners are to consider whether it is absolutely necessary to observe gender at all, as opposed to simply asking respondents to provide a self-report. In addition, if collecting respondent gender by interviewer observation is a firm's preferred method, survey designers should consider what placement would most reduce errors.

In this dissertation, to this point, I have shown that: (1) survey data collection firms often use interviewer observations to collect respondent gender; (2) there are notable errors in these observations which are correlated with the length of time interviewers are exposed to the voice. What is left to explore is the consequence of these errors. The next chapter shows how final survey estimates and inferences about findings are affected by the misclassification of respondent gender in interviewer observations.

4. Understanding the Consequences of Observational Gender Misclassification on Survey Estimates

Survey methodologists focus on study design decisions, questionnaire wording, data collection procedures, weighting procedures, and many other processes that affect sources of error in final survey estimates. However, analysts often focus on how survey results change due to various design decisions. And, what often matters is the bottom line: "What does the estimate of interest tell us about a population?" For example, what is President Obama's approval rating among Americans and how is it different between men and women?; What is the incidence of lung cancer screening among women?; How many men would purchase Gillette razors over Schick brand razors? These examples reference differences in estimates by gender – one of the most commonly used variables used when describing behaviors, attitudes, knowledge, or opinions. However, no known research has addressed how final survey estimates may be affected by the methodology used to collect respondent gender. Does President Obama's approval rating look different among men and women using a respondent's report of their gender as opposed to an interviewer observation of their gender? If so, how gender is collected in a survey might change analysts' conclusions, recommendations, or inferences made. Focusing on telephone data collection methods, this chapter provides, to my knowledge, the first evaluation of the impact and potential consequences of measurement error in respondent gender data.

According to the study described in Chapter 2, 68% of survey organizations that perform telephone data collection determine respondent gender by observation – either a pure observation (30% of organizations) or *ask-assisted* (38%), that is, aided by an

instruction for interviewers to ask only if the respondent's gender is not obvious. Fifteen percent of firms ask respondents whether they are a male or female. The survey found that the remaining 17% utilize multiple methods or some other method.

If there are errors in interviewer observations of respondent gender, then survey estimates by gender could be affected. For example, as mentioned in Chapter 1, the Health Information National Trends Survey (HINTS) was a national telephone study commissioned by The National Cancer Institute from 2003-2007, which provided data about access to cancer information, perceptions of cancer risks, and patterns of health care needs. Using the ask-assisted method of gender data collection, interviewers observed the respondents' gender early in the survey. Only when it was "not obvious" were interviewers instructed to ask whether they were speaking with a man or woman. If, in fact, there is some error in the HINTS gender data, conclusions derived from observed differences between men and women may be flawed. For example, McQueen et. al (2006) use the 2002-2003 HINTS data to better study screening prevalence for colorectal cancer among males and females. They find some significant differences in knowledge, opinions, communication, and testing practices between the two genders, noting the importance of identifying gender-specific strategies to increase colorectal cancer screening. Assuming the researchers used the only known gender identifier in the dataset, the ask-assisted interviewer observation, to distinguish responses by gender, their estimates are likely to include some measurement error, which would affect their final estimates, and, perhaps, their conclusions and policy recommendations. The extent of measurement error in the gender variable in the HINTS survey is unknown, however, and the effects of this measurement error are also unknown.

Why would I suspect that observational methods of gender identification would produce measurement error? As discussed in the introduction to this thesis, there is empirical evidence that gender coding by observation leads to misclassification. That preliminary research suggested some differences in survey outcomes when using the interviewer report of gender compared with the observation. Chapter 3 also documented errors in interviewer observations of respondent gender in an experimental setting. There, the misclassification was found to be particularly strong for short exposures to the voice. This finding implies that interviewers observing gender early in the survey are particularly likely to induce errors. The question addressed in this chapter is: How much does this misclassification affect survey estimates?

The misclassification found in the experiment in Chapter 3 was not random, but systematic to certain situational and speaker level characteristics. As Lessler and Kalsbeek (1992) describe the errors associated with various forms of data collection, they note that direct observations are sometimes flawed. Interviewers may be faced with confusing pitch signals, unexpected or non-stereotypical behaviors, or short exposure lengths, creating a non-representative depiction of the respondent's gender. Lessler and Kalsbeek go on to explain that such errors can result in an observer bias – possibly influencing final estimates. If this bias assessment holds for the general survey telephone setting, survey researchers may need either to use interviewer observations more cautiously or intervene (for example, through retraining interviewers).

Given the prevalent use of observational methods to collect respondent gender, here I seek to understand the consequences of error in observational reports of gender on survey estimates. That is, to what extent is the potential variation in the quality of gender

classification problematic for survey practitioners? Using secondary analysis of multiple datasets, this paper addresses the following question:

1. What differences in survey estimates are obtained and what is the bias when using interviewer observations of respondent gender for analysis?
2. Would different conclusions be made when using true-values of gender versus interviewer observations of respondent gender to identify statistical differences between male and female survey estimates?

4.1. Data and Methods

Using data from 28 independent public opinion telephone surveys collected by The Marist College Institute for Public Opinion (MIPO), I conducted secondary data analysis to evaluate measurement error induced by inaccurate interviewer observations of respondents' gender, and estimate the impact of this error on important survey results.

4.1.1. Data Description

The analyses in this paper rely on 28 different studies conducted by MIPO from September 2008 to February 2010. Pooling across these different studies: (1) strengthened the reliability of the findings by eliminating the possibility that findings were unique to the observational error obtained in one, particular project; and (2) increased efficiency through a larger case base. The surveys included in this research are all of the public studies (not privately funded or commissioned by an external client) from the time that MIPO began including both interviewer observations of respondent gender as well as respondent self-reports (2008) through the time that the data was pooled (2010).

All studies included here were conducted at MIPO's centralized telephone facility in Poughkeepsie, New York. Adult residents (aged 18 years and older) from three sample frames are represented in these data: 14 studies were conducted among residents of the contiguous United States; 7 studies were conducted only in New York State; and 7 included studies were conducted in New York City.

Across all surveys, there were 26,221 respondents: the smallest had a sample size of 644 and the largest a sample size of 1176. The average sample size across the 28 projects was 891. The average interview length was approximately 9.5 minutes; the shortest survey being 5.5 minutes in administration length; the longest running, on average, 15 minutes. Only data collected from random-digit-dial (RDD) landline telephone frames were included; data from the cell phone frames were excluded from analysis. The reason only landline surveys were included in this research was because not enough cell phone cases, with both necessary gender data points, were available at the time the data was compiled.

As shown in Table 4.01, the majority of respondents in the pooled dataset were over the age of 45 and white. Gender proportions were more evenly distributed (44% male; 56% female). A total of 475 unique[20] interviewers completed at least one of the surveys included in this study. All MIPO interviewers are undergraduate students at Marist College and undergo extensive telephone survey training. Although interviewers are not specifically trained in distinguishing gender voices and pitch tones, they are told of the importance of their gender observations. Table 4.01 also presents interviewer

[20] I verified with MIPO's Survey Operations Manager that all interviewers are assigned a unique ID.

demographic characteristics - including their race, gender, and experience. By virtue of being undergraduate students, all interviewers were between the ages of 18 and 23.

Table 4.01

Unweighted Respondent and Interviewer Demographics

	Respondent Demographics	**Interviewer Demographics**
Age		
Under 45	26%	100%
Over 45	74%	0%
Race		
White	78%	81%
African-American	10%	8%
Other race	8%	11%
Refused	4%	n/a
Gender		
Male	44%	29%
Female	56%	71%
Mean Experience *(measured by number of completed interviews)*	n/a	54

Each study included its own set of questions, although most of MIPO's questionnaires address topics surrounding politics, current events, and social attitudes. As discussed below, some questions appeared in nearly all of the studies (e.g., demographic questions); while others were unique to a specific study. Thus, after pooling, sample sizes varied from question to question (Appendix C presents the sample sizes for the questions included in the analysis).

Aside from having respondent and interviewer characteristics, the unique feature of this dataset is the availability of both an interviewer observation of respondents' gender (*observe*) as well as a respondent self-reported gender (*reported*). Across the 26,221 pooled observations, 1,285 observations were omitted from analysis. This left 24,936 cases for analysis. Respondents refusing to provide a verbal response to their

gender or, for an unknown reason the gender data was missing from the dataset, accounted for 525 of the cases. A total of 104 cases were removed because interviewers could not make even a guess of the respondent's gender (average of 3 such responses across the included studies – minimum of 0; maximum of 7). Given both an interviewer observation and respondent report were necessary to assess the measurement error, observations with missing data for either of these data points (described above) were excluded from the analysis. The remaining 636 cases were omitted because they were a cell phone interview (recall only landline interviews are analyzed in this research).

The placement of both items was consistent across all studies: MIPO interviewers are asked to observe the respondent's gender immediately after the survey introduction. Interviewers are trained to make a best guess and avoid, at all costs, using the result code 'could not make guess.' Respondents are asked to state their gender at the very end of the questionnaire. MIPO's household selection procedure asks to speak with the youngest adult male if more than one adult lives in the household. In such instances interviewers repeat the gender observation once the respondent comes to the phone. Below depicts the survey introduction and questionnaire flow used in all of the included studies.

Q1. Hello. My name is <name>. I'm calling from Marist College. We're talking to people in your community and collecting opinions about issues facing residents. Are you 18 years of age or older?

Yes
No – *Interviewer ask to speak to another member of the household who is 18 and restate introduction*

Q2. *Interviewer*: Record gender of person on the phone by observation only:

Male
Female
Could not make a guess

Q3. How many adults, aged 18 or older, currently live in your household?

1 – *Skip to Survey Content*
2
3 or more

Q4. May I please speak with the youngest male who is currently at home?

Yes – *Once new respondent comes to the phone, interviewer re-states survey introduction, makes another observation of respondent gender, then continues with full survey*
No – *Continue to full survey with person on the phone*

<SURVEY CONTENT>

Last Q. Are you:
Male

4.1.2. Outcome Variables

The pooled dataset of 28 studies contained 998 variables. The 998 variables were not all unique measures as many were recodes of other variables in the dataset (e.g., continuous age variable recoded into categories; President Obama's approval rating coded dichotomously and categorically). However, all were retained for analysis since they were not identical variables (e.g., same question with the same recording/response options). Given MIPO names variables with the same measure consistently across studies, duplicity of identical variables did not occur.

As mentioned, some questions were only asked in one study while others, such as demographic questions, appeared in nearly all studies. All questions, regardless of sample size, were retained to allow for the analysis of variables with both large and small sample sizes. While I do not separate the findings by sample size, the importance of this issue is addressed in the results section.

Seventy-five variables were initially selected for analysis using simple random sampling. The random selection was conducted by listing each variable in Excel and assigning a random number. The file was then sorted in numeric order and the first 75 variables were selected. Twelve of the variables were either a recode of an already selected variable or one that did not contain respondent data (e.g., a transition statement such as "Switching topics…) and replaced with the next variable in the file. Only 50 of the 75 variables were used in the analysis (again, selected at random from the list of 75

variables).[21] Appendix C displays the question wording, and sample size of the selected variables included in this analysis. They represent a variety of topics and types – both demographic and attitudinal.

4.1.3. Nature of Gender Misclassification in Dataset[22]

Assuming the respondent reported gender is the true value, across all of the 28 studies, on average, 8.3% of the gender observations made by the interviewers were incorrect. The gender misclassification rate ranged from study to study: a low of 4.49% in one study, to a high of 13.85% in another. Table 4.02 shows the misclassification by project along with other study information (e.g., dates of data collection, number of interviewers, and percent of African-American women in the sample since this is the group most likely to be misclassified by interviewers). In terms of causes of the wide variability in misclassification, the information suggests no systematic patterns but, it is notable that the two studies with the largest error rate are also those with the two smallest sample size and number of interviewers. Also, as previously mentioned, the introduction sponsor, and characterization of the study is the same across all of the projects. When

---

[21] Initially, I planned to conduct the analysis on 75 variables; however, once the dataset was subset, I decided to reduce the number of analytic outcomes to 50 for clearer presentation of the findings.

[22] Since the data source is the same, the data presented in section 4.1.3 mirrors findings included in McCulloch et al. (2010). Please note, however, that some figures may show very small differences (e.g., in Table 4.04, the proportion of African-American females who were observed male was reported as 18.3, while this work shows it as 18.1). This is due to the exclusion of cases in this work with item missing data for the respondent sex report and/or the interviewer observation.

asked what the survey is about, MIPO interviewers are trained to not disclose the topic of the survey, and instead, use generic responses such as “it’s about issues facing residents.” Thus, there is little reason to believe that the topic of the survey would have elicited more conversation which could have improved the quality of the interviewer gender observation.

Table 4.02

Description of Each Study in Pooled Data Including their n, Dates of Data Collection, Number of Interviews, Proportion African-American Women, and Overall Gender Misclassification Rate

Population	Number of Observations	Month/Year of Data Collection	Number of Interviewers	Unweighted Percent African-American Women	Percent of Observational Gender Misclassification
United States	1,114	06/09	54	4.31	4.49
New York State	697	01/10	65	5.88	4.59
New York City	823	10/09	91	13.37	5.47
New York City	714	09/09	89	12.75	5.60
New York State	746	11/09	86	4.29	5.63
United States	786	09/08	63	4.33	5.73
United States	801	11/08	32	4.37	5.74
New York State	709	09/09	81	4.09	5.92
United States	923	10/09	79	2.93	6.18
United States	1,030	02/10	87	3.88	6.60
New York City	914	10/09	82	12.80	6.78
United States	956	12/09	86	2.82	7.95
United States	989	09/08	63	4.04	7.99
United States	862	08/09	49	3.36	8.00
New York State	975	06/09	64	5.64	8.00
New York City	915	11/08	67	13.99	8.09
United States	1,176	04/09	101	2.81	8.33
New York City	812	06/09	65	12.32	8.62
United States	962	11/08	41	5.09	8.73
New York City	963	12/08	52	4.05	10.38
New York City	699	05/09	70	9.73	10.44
New York State	1,021	02/09	66	5.29	10.58
New York City	801	02/09	73	11.61	11.11
New York City	1,116	03/09	81	3.49	11.11
United States	1,099	04/09	70	3.28	11.46
New York State	1,003	04/09	65	2.99	11.67
New York State	644	10/08	43	6.21	12.89
United States	686	10/08	33	4.66	13.85

As shown in Table 4.03, there was significant differential measurement error between male and female respondents. For the females, 12.7% were incorrectly classified as male by interviewers compared with 2.6% of male respondents who were

misclassified. As a consequence, when we think of the pooled data and use interviewer observations as analytic variables, approximately 14% of all respondents judged to be males by interviewers are, in fact, females.[23]

Table 4.03

Respondent Reported Gender vs. Interviewer Observed Gender

MIPO Data	Respondent Reported Male	Respondent Reported Female	Total
Observed Male	97.4%	12.7%	49.5%
Observed Female	2.6%	87.3%	50.2%
n	10,878	14,058	24,936

A bivariate table of incorrect interviewer gender observations revealed some systematic differences across various racial groups of respondents (Table 4.04). African-Americans, especially, were most likely to have a wrong gender assessment as interviewers incorrectly categorized 12.6% of all respondents in this subgroup.

Table 4.04

Errors in Interviewer Gender Observations Across Racial Groups

MIPO Data	White	African American	Hispanic	Asian	Other	Refused Race	Total
Observed Incorrect	7.9%	12.6%	7.8%	8.0%	7.3%	8.3%	8.3%
n	19,379	2,360	1,077	537	666	907	24,936

Given the large proportion of misclassification within African-Americans and women, I break the error rates down by true gender for African-American respondents.

[23] There are a total of 12,378 observed male respondents. There are a total of 1,786 true females but observed males in the dataset.

In terms of the direction of these errors, Table 4.05 shows that 18.3% of African-American women were miscoded as men, while only 2.9% of African-American men were perceived to be women by the interviewers. This result suggests the need to investigate how estimates of substantive characteristics of African-American women, such as voting behavior, may be biased.

Table 4.05

Direction of Errors in Interviewer Gender Observations among African Americans

MIPO Data		Respondent African American Male	Respondent African American Female	Total
Observed Male		97.1%	18.3%	47.1%
Observed Female		2.9%	81.8%	52.9%
	n	864	1,496	2,360

4.2. Analysis Methods

All of the 50 variables under analysis were categorical with a plurality having 4 valid response options. The extreme categories of variables with more than two options typically have fewer responses than do the middle categories. Thus, misclassification in interviewer observations of respondent gender would likely move a larger fraction of the responses in these extreme categories. Because of this, I created binary outcome variables, focusing on extreme answers. That is, regardless of the number of categories in the original variable – values of 1 were retained as a 1 in the binary outcome; all other values were assigned a 0. For example, the question "How concerned are you that you and your family may need to turn to food assistance?" had four original response options: very concerned, concerned, not very concerned, not at all concerned. This question was recoded into a binary variable: 1 being very concerned, all other responses becoming a 0.

In another example, the question: "Would you rate the job Senator Hillary Clinton has done in office as excellent, good, fair, or poor?" was recoded into a binary variable with 1 being all 'excellent' responses, 0's being the good, fair, and poor responses. The decision to analyze binary outcomes also eases comparison across the 50 variables since the number of response options varied across variables. I recognize that the decision to create binary outcomes based on one response category results in the possibility of missing notable differences that may occur in other categories. However, for the purpose of this chapter, calculating bias in estimates for even one set of responses, illustrates the possible effect of misclassification in interviewer observations. All analysis was conducted using STATA 12.

4.2.1. Biases in Estimates of y for Each Gender Group

With the goal of documenting whether interviewer misclassification of respondent gender affects survey estimates, the first set of analyses looks at the differences in estimates by gender that are obtained when analyzed with the interviewer observation versus the respondent report. Differences in these simple, bivariate findings substantiated the need for further investigation of the overall bias across the binary variables.

Similar to the approach used by Eckman and Kreuter (2012) to examine undercoverage bias in traditional housing unit listings, bias and relative bias assessments were calculated to determine the size of the effect of gender misclassification. Figure 4.01 visualizes the data at hand in a Venn diagram, focusing on estimates for females only. The left side circle represents the true females, and the right side circle represents all respondents observed by the interviewer to be female. Segment A thus represents those females that are a true female but were observed to be a male and would be left out

of analyses if the interviewer observation of gender is used as an analysis variable. The segment labeled B indicates those that are a true female and observed female. Thus, $A \cup B$ represents the means for true (correct) females. The segment C on the other hand represents those that were observed female but are a true male. Thus, $B \cup C$ represents all cases that would be labeled as female in an analysis that uses the interviewer's gender assessments. The bias in the resulting estimates can thus be calculated by taking the difference between the average for the outcome variable of interest (y) for the true females ($A \cup B$) minus the average of the outcome variable for those assessed as being female ($B \cup C$):

$$bias(\bar{y}) = \bar{y}_{BC} - \bar{y}_{AB} \qquad (1)$$

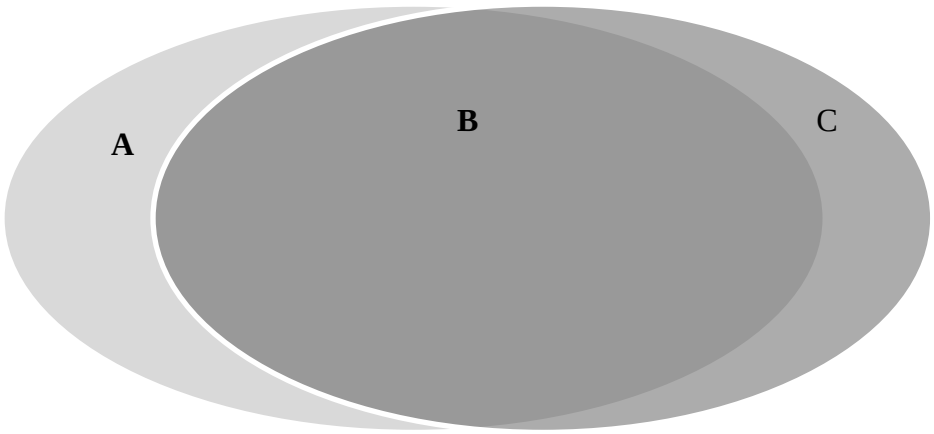


Figure 4.01: Venn Diagram of Gender Assignment for True or Observed Females

Table 4.06 provides another way of looking at the different sets and the resulting categorization used for the estimation of bias in y.

Table 4.06

Example Gender Classifications for Females and Resulting Venn Diagram Category

Self-Reported Gender	**Interviewer Gender Observation**	**Category in the Venn Diagram**
Female	Female	B
Female	Male	A
Male	Female	C

"The absolute value of a bias does not provide much information on the impact of the bias on estimates" (Bose, 2001, p.7). Thus, to provide a measure of the magnitude of the bias, that is, to evaluate the bias across different survey outcomes, I calculate and present the relative bias. This is the ratio of the bias and the mean estimate using the true gender:

$$relbias(\bar{y}) = \frac{\bar{y}_{BC} - \bar{y}_{AB}}{\bar{y}_{AB}} \qquad (2)$$

I expect the relative bias to be greater for questions that show more gender differences. Given this, I display the results of the relative bias separately for attitudinal, behavioral, and demographic questions. For example, often males and females hold different attitudes and opinions about various issues (e.g., beliefs about whether restaurants should show their calorie counts per serving on menus). Given misclassification, differences in male and female estimates may be exacerbated when using interviewer observations. Similarly, I expect behavioral questions (e.g., voting behaviors) to be affected by misclassification. Demographic questions, such as race and age, may also be affected since females (and subgroups within females such as African-American and older females) are more likely to be misclassified. This results in more

older males and African-American males in the data when observational data is used. Thus, the 50 binary outcomes were categorized into one of these three question types.

4.2.2. Biases in Estimated Differences between Gender Groups

A change in distribution is one thing, but ultimately what may be important is whether different conclusion would be drawn from the data. Thus, the second set of analyses evaluates whether directionality and significance of differences between males and females on survey estimates change as a result of the misclassification. Often tests of proportions are used to examine differences between male and females on some outcome variable. Since we have two gender measures (one using the respondent report and the other using the interviewer observation), we can do two tests to determine if they reach the same conclusion.

In the first test I obtain the difference between male and females and its corresponding 95% confidence interval (CI) using the respondent report; the second obtains the difference between observed males and females and its corresponding CI using the interviewer observations of respondent gender. The results of these tests are plotted in several graphs. Each of the fifty survey variable has two data points on the graphs (one showing the difference between males and females using report; the other using observation). In these graphs, two assessments can be made: (1) the difference between the two point estimates, and (2), whether the different point estimates are significantly different from 0. Because data are pooled across different surveys no selection weights are used, and because nonresponse weights often include gender as a stratifying variable, those are not used either in these tests.

To structures this analysis, I divided the set of dichotomous variables into two groups: one where there is a statistical difference between males and females; the other where no difference is observed. This determination is made using the true-value (respondent self-report) of gender in tests for proportions. I only expect differences in the results between classifications by self-reports vs. observation to occur in those variables that show differences between males and females when using self-reports of gender. If males and females do not differ in their attitudes or behavior, then a misclassification should not matter. However, for sake of completeness the second set of variables is displayed here as well.

4.3. Results

The results are presented in two sections to address each of the research questions. First, I present the difference in estimates obtained from the two gender reports and whether there is a bias in final male and female estimates using the respondent report verss the interviewer observation. Second, I discuss the statistically significant conclusions drawn from, first, a t-test with the true –value of gender, and second, a t-test with the interviewer observation of respondents' gender.

4.3.1. Is there a Difference and Bias in Estimates When Using Gender Observations?

Before presenting the overall results from all 50 binary estimates, Table 4.07 gives an example for the difference in the full response distribution when using one or the other gender variable. Of all males, 12.27% report being very concerned that they may need to turn to food assistance using the respondent report. This number increases by

approximately two percentage points (14.04%) when using the observational data to determine gender. For females, there are also shifts in estimates between the two gender reports. Note, that the estimates show the largest differences at the extremes (hence, motivating the decision to analyze the binary outcomes).

Table 4.07

Concern for Food Assistance When Using Various Gender Reports by Gender

	How concerned are you that you and your family may need to turn to food assistance?			
	Distribution among men when using respondent gender report (n=693)	Distribution among men when using interviewer gender observation (n=762)	Distribution among women when using respondent gender report (n=1020)	Distribution among women when using interviewer gender observation (n=951)
Very concerned	12.27%	14.04%	17.25%	16.19%
Somewhat concerned	14.00%	14.30%	16.18%	16.09%
Not very concerned	22.37%	22.83%	21.47%	21.03%
Not at all concerned	51.37%	48.82%	44.41%	45.95%
Unsure	0%	0%	.69%	.74%

In another example (Table 4.08), a question that asked NYC residents whether they think a law that requires restaurants to show calorie counts for menu item should be displayed on an item or serving level, a similar effect is observed: 31.27% of true males report that restaurants should show their calorie counts on a per serving basis; whereas this estimate changes to 34.15% when using the interviewer observed males. For true females, 43.36% say restaurants should show their calorie counts per serving; when using the observed females, this estimate changed to 41.83%.

Table 4.08

Attitudes Toward Calorie Counts on Menus When Using Various Gender Reports by Gender

	Do you think a law that requires restaurants to show calorie counts for items on their menu should show nutritional information on a per serving basis or do you think restaurants should show calorie counts for the entire item, regardless of the number of servings? (n=1097)			
	Distribution among men when using respondent gender report (n=502)	Distribution among men when using interviewer gender observation (n=571)	Distribution among women when using respondent gender report (n=595)	Distribution among women when using interviewer gender observation (n=526)
Per serving	31.27% (n=157)	34.15% (n=195)	43.36 (n=258)	41.83% (n=220)
For the entire menu item	45.82 (n=230)	43.08 (n=246)	39.83% (n=237)	42.02 (n=221)
Unsure	22.91 (n=115)	22.77 (n=130)	16.81% (n=100)	16.16 (n=85)

Using the 50 randomly selected binary variables, Figure 4.02 presents the percentage point difference between the proportions obtained using the respondent self-reported gender and those obtained when using the interviewer observation for males and females, respectively. This descriptive graph orders all of the differences in outcome measures in ascending order. Thus, it is important to note that the survey questions on the y-axis are ordered differently for males and females. The rationale behind ordering the graphs in this way was to easily identify a pattern and the questions which appear most affected by gender misclassification and show all of the question types together. Differences in the binary outcomes are small: ranging from -2.3 to 2.9 (average absolute value of 2.9 percentage points) for males. For females, differences in proportions for females were slightly smaller, ranging from 1.9 to 1.8 (average absolute value of 1.9 percentage points) for females. There are three questions in which there is no difference between the proportions obtained in both gender measures (dot is at 0) for males. Those questions are variable 10 (whether they favor legalizing same-sex marriage in New York

State), variable 23 (whether they think economic conditions were inherited by Obama), and variable 44 (whether the respondent is a Latino). Looking at the most extreme case (also presented in Table 4.08) for males, the top line (variable 7), the male proportion for the survey question: "Do you think a law that requires restaurants to show calorie counts for items on their menu should (value 1) show nutritional information on a per serving basis or (value 0) do you think restaurants should show calorie counts for the entire item, regardless of the number of servings?" is a 31.3% when obtained from the self-report. However, the male proportion increases to 34.2% when using the interviewer observation. The dot represents the difference: a 2.9 percentage point difference in the estimates for this binary outcome question when using the two gender measures. Looking at the same question for females, 43.4% females when self-identified and 41.8% using observed females are proponents of restaurants calorie counts per serving rather than showing the calorie counts for the entire menu item. This results in a difference of 1.6 percentage points. For females, there are four questions which show no difference in the estimates. While these differences may not be large, they do show that misclassification can lead to some movement in male and female survey estimates.

Comparing the male and female line patterns, although the differences are subtle, it shows that males have slightly greater differences between the estimates obtained when using the self-report rather than the interviewer gender observation. This is evident by the female line being slightly steeper and the dots being less dispersed from the center 0 line. The male estimates are affected by gender misclassification more than the females. Recall, this reflects the fact that females are more likely to be observed male than are males likely to be wrongly coded female. Thus, there is an increased amount of "wrong"

data in the true male estimates.

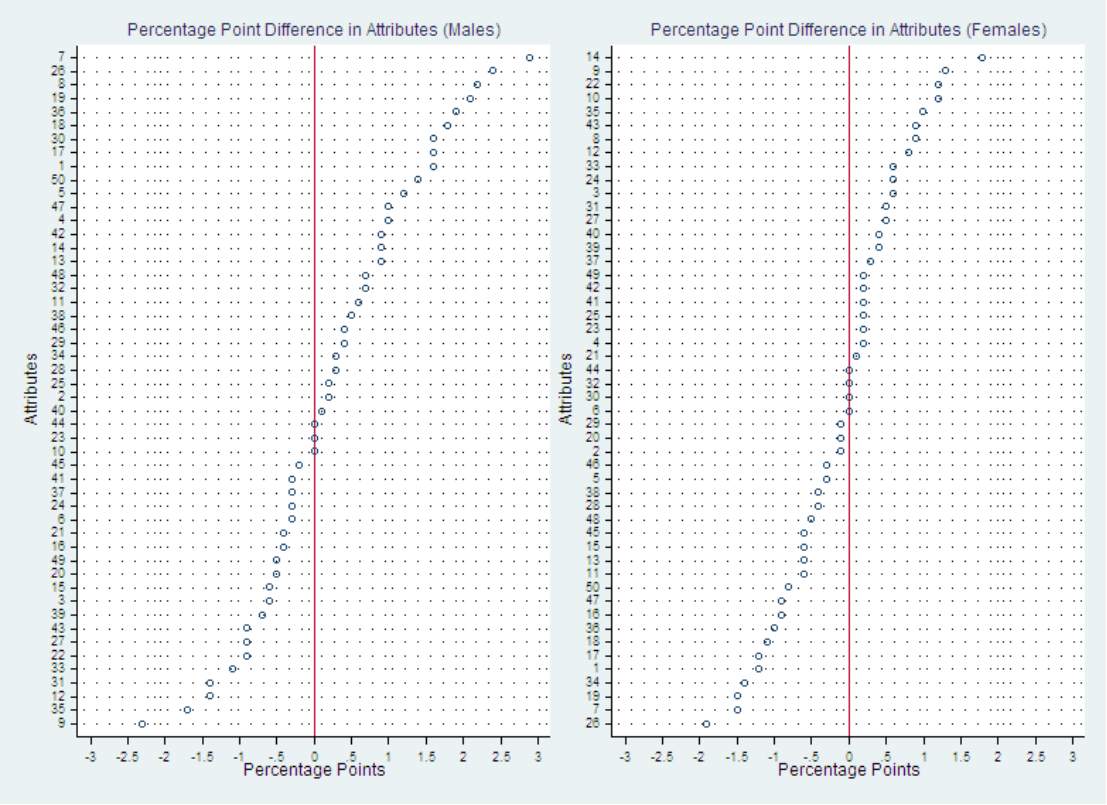


Figure 4.02. Percentage Point Difference in Proportions Using Gender Observation vs. Respondent Self-Report

The relative magnitude of these differences is explored next. Figure 4.03 presents the *relative* bias in the means for the 30 attitudinal questions (descriptions of the variables are provided in a table after the graph). The result for males is presented on the left side; females are on the right. Here, we see a wide range of relative bias: from -26% to +54%, although the average relative bias for males and females is 4.26% and -1.56%, respectively. Overall, the misclassification of respondent's gender in interviewer observations affects the male estimates more than the females for attitudinal variables – indicated by the dots being slightly more dispersed from the 0 line. The relative bias is greater for the males than females. Males are less likely to be misclassified, thus,

resulting in less wrong female data. Looking at the first row, variable 30 is a measure of Michele Obama's approval rating (1= Rates Michele Obama's role as excellent; 0= Rates Michele Obama's role as good, fair, or poor). This is a variable where one might hypothesize gender differences exist – females being more likely than males to say Michele Obama's role has been excellent. The left side graph shows that the mean number of wrongly observed males who rate Michele Obama's role in office as excellent is approximately 9% higher, relative to the mean from all known, self-reported males. Considering the overall misclassification is greater among females (more likely to be classified male than men are to be observed female), this finding is consistent. That is, if females are more likely to approve of Michele Obama and be wrongly assigned a male, there would be more male observations with female responses when using observational data. Conversely, for females (right side graph), this variable shows little relative bias (-.1%) – notably less than was observed for males. This means that Michelle Obama's approval rating is 1% lower, relative to the mean from all, self-reported female respondents. Another example, variable 15 (measuring whether or not respondents think the U.S. economy is getting better or worse), shows gender misclassification causes little relative bias for both males and females, and in the same direction (mean being approximately 2% lower, relative to that of known males and females).

Bias induced by interviewer misclassification of respondent gender also appears to be affected by sample sizes. For example, looking at variable 26, a question assessing whether or not people think schools in their community are prepared to deal with a disaster situation, shows the highest relative bias across all of the attitudinal variables. This is not a question where, intuitively, one might expect large differences between male

and female opinions. For this question, the number of males who were wrongly observed female was 12; whereas the number of females who were wrongly observed male was 57. The proportion of males who think schools are very prepared to deal with a disaster is 8.33% when looking at the cases that were wrongly observed female; however, it is 4.56% among true males where there was no observational error. For females, 22.8% of the cases who were wrongly classified male say schools are prepared; however, only 5.58% of true females share this view. This large difference between the females observed female and those that were classified male, results in the male estimate to be approximately 50% higher, relative to the mean from all known, self-reported males. This implies that the small sample sizes in the wrongly observed cells, makes the estimates more susceptible to error.

Figures 4.04 and 4.05 shows the relative bias for other types of variables. Overall, the average relative bias is slightly smaller for behavioral and demographic variables as it was for the attitudinal outcomes. The range is significantly less (-9 to +11) and the average relative bias for both males and females is slightly less. For the behavioral questions, the male average relative bias is -1.65; for females, it is .85. For the demographic questions, the male average relative bias is -2.16; for females, it is -1.21. This implies that the miscoding of respondent gender by observation has a substantial effect on final survey estimates if there are large, inherent differences between males and females or if the sample sizes are small.

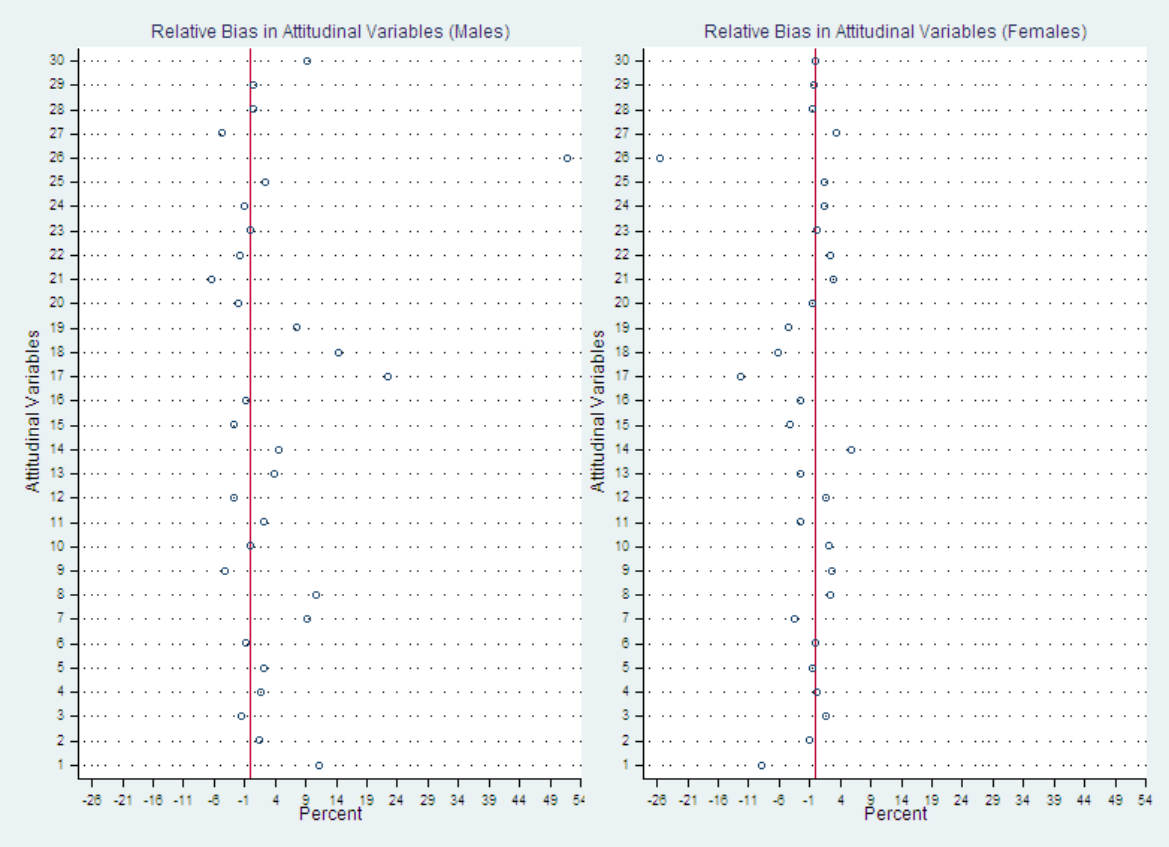


Figure 4.03 Relative Bias in Attitudinal Variables

Variable	Description of ‘1’ Value	n
30	Rates Michele Obama's role as excellent	1091
29	Thinks there should be a law banning texting while driving	923
28	Says Tiger Woods should address the recent events in the news privately	568
27	Says they have had a vacation considered to be a disaster	850
26	Thinks schools in community are very prepared to deal with a disaster	850
25	Is very concerned about getting H1N1 flu virus	1758
24	Thinks that things in the country are going in the right direction	9523
23	Thinks economic conditions were inherited by Obama	6366
22	Says Obama has met their expectations	1029
21	Thinks the hardest time to raise a child is from infancy through walking	1121
20	Has a favorable impression of Harold Ford	307
19	Always worries that family income will not be enough	635
18	Very concerned that family may need to turn to food assistance	1713
17	Thinks it’s very likely people will be confused how to cast their ballots in 2008 election	686
16	Thinks the number of US troops in Afghanistan should be increased	912
15	Thinks US economy is getting better	1070
14	Thinks 5 and under is age to start talking to kids about money	1059
13	Approves of the job Republicans in Congress are doing	968
12	Says things in New York City are going in the right direction	5642
11	Says Bloomberg has handled education best as mayor	701
10	Favors legalizing same-sex marriage in New York State	739
9	Approves of how Mayor Bloomberg is handling the city's budget	2168
8	Rates Hillary Clintons job in office as excellent	641
7	Thinks restaurants should be required to show calorie counts per serving	1097
6	Says Obama should cut taxes even if it means more debt	1935
5	Rates Senator Schumer's approval rating as excellent or good	5134
4	Prefers next president deals with health care crisis over cutting taxes	1936
3	Approves of Congress giving federal loans to US automakers	958
2	Rates Mayor Bloomberg's job in office as excellent	800
1	Thinks NYC bus and subway system is getting better	775

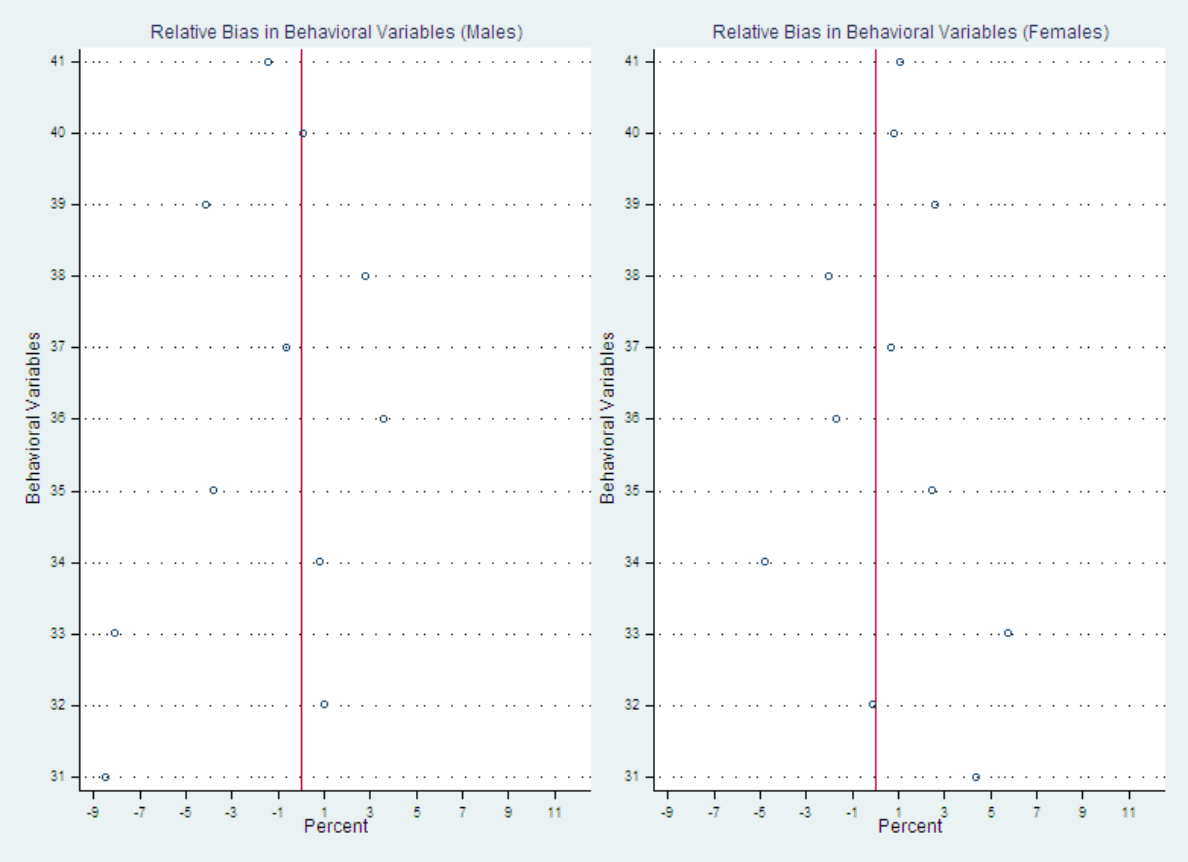


Figure 4.04. Relative Bias in Behavioral Variables

Variable	Description of '1' Value	n
41	Went to live performances such as plays often	1104
40	Plans to watch a great deal of Palin-Biden debate	1744
39	Plans to spend tax refund rather than pay bills or save	680
38	Plans to watch a great deal of the winter Olympics	1023
37	Plans to vote for Paterson in 2010 Governor race	4966
36	Plans to vote for Cuomo in the 2010 NYS Gov election	4410
35	Plans to vote for McCain-Palin in 2008 election	3706
34	Plans on watching all of the Super Bowl	1019
33	Eats dinner at home four or fewer days a week	579
32	Is very likely to complete the 2010 Census form	1008
31	Has done something to reduce spending money	1116

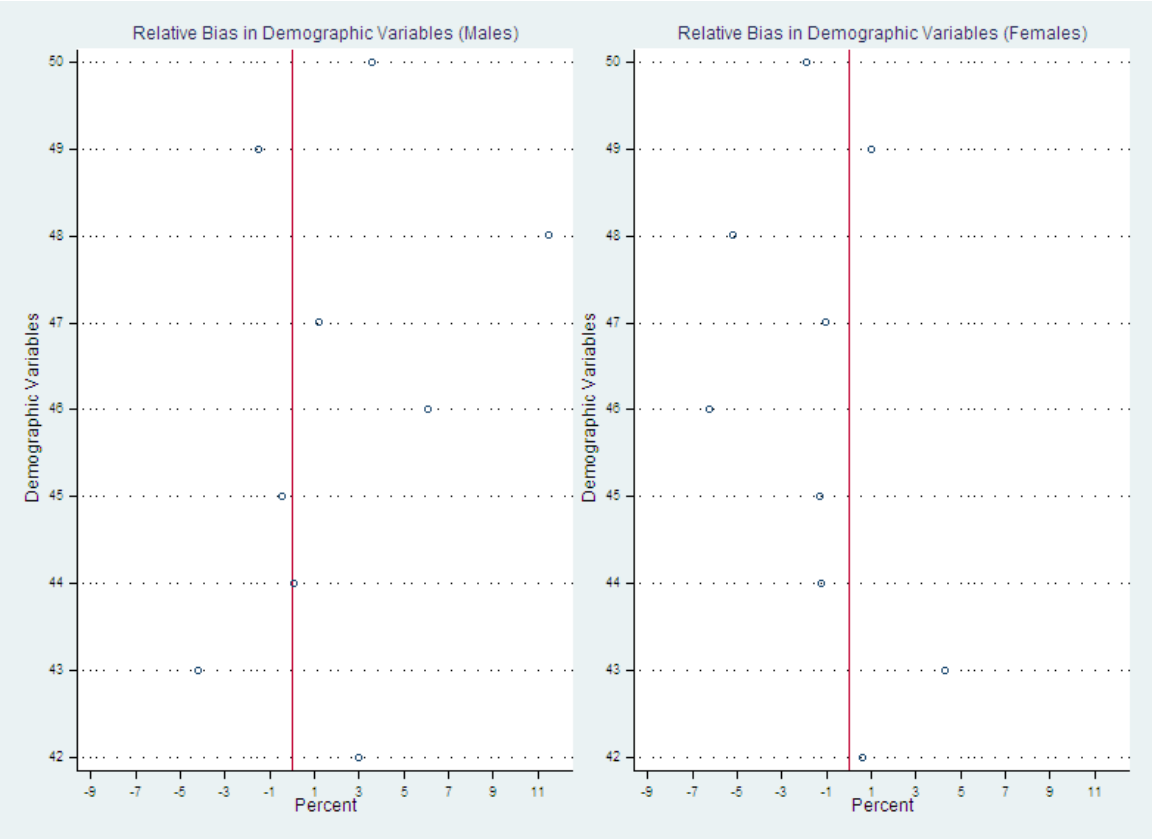


Figure 4.05. Relative Bias in Demographic Variables

Variable	Description of '1' Value	n
50	Is a Protestant	22217
49	Is under 45 years old	1744
48	Has an annual income of less than $15000	1560
47	Is a household with health insurance	1166
46	Has a Twitter account	1100
45	Is employed	2860
44	Is a Latino	22885
43	Considers themselves an environmentalist	1098
42	Reports being a Democrat	7962

4.3.2. Do Statistically-Driven Conclusions Change Depending on Which Gender Data is Used?

As discussed in the analysis section, I compared the results from tests for proportions using (one using the respondent self gender report and the other using the interviewer observation) to identify whether gender interpretations might be different when using the respondent report compared with the interviewer observation. The resulting z-tests determine whether there is a significant difference between the male and female mean estimates. A significant p-value for Ha indicates a rejection of the null hypothesis (there is a significant gender difference).

Using one outcome variable, whether or not respondents are very concerned about needing to turn to food assistance (n=1713), by way of example, in Table 4.09, I show some of the output of the test of proportions when using the two gender measures.[24] Comparing the results, it is possible that a researcher is could draw different conclusions if using one gender measure vs. the other. In the top half of the table, the test uses the interviewer true-value of gender and suggests that there is a significant difference between males and females when it comes to their perception of needing to seek food assistance (p=.0048). However, using the interviewer observation of respondent's gender (the bottom half of the table), a researcher could accept the null hypothesis and conclude that there are not significant differences between males and females (p=.2182) when it

[24] These tests were run using the prtest command in STATA (comparison of proportions, producing a z statistic). By way of comparison, I also ran tests using the ttest command (comparing means, producing a t statistic). The results between the two tests are nearly identical.

comes to this variable. This shows that it is possible for a researcher to draw different conclusions, depending on the gender collection method, using p-values to inform that judgment. However, a comparison of the confidence intervals of the differences shows they overlap, suggesting there is not a statistical difference between the male/female estimates by method of collecting respondent gender. Graphing this information provides an efficient way of seeing (1) differences in the point estimates; and (2) their differences from the 0 line.

Table 4.09.

Gender Differences for Concern for Needing to Turn to Food Assistance by Two Measures

Very concerned that family may need to turn to food assistance			
	Proportions	95% Confidence Interval	Two-tailed p-value
Self-reported males	12.3%	11.6%-16.5%	
Self-reported females	17.3%	13.9%-18.5%	
DIFFERENCE	-5.0%	-8.4%-1.6%	0.0048
Very concerned that family may need to turn to food assistance			
	Proportions	95% Confidence Interval	Two-tailed p-value
Observed males	14.0%	11.6%-16.5%	
Observed females	16.2%	13.9%-18.5%	
DIFFERENCE	-2.2%	-5.6%-1.2%	0.2182

Figure 4.06 graphs all of the variables for which there is a statistically significant difference between self-reported males and females. That is, they are the 24 variables that, according to a test of proportions (z-test), there is a significant gender difference in response. Each of the 24 variables has two lines on the graph (e.g., attribute47.1 and attribute47.2). Those with a .2 (also denoted by the green line and triangle) represents the

difference between male and female proportions using the interviewer gender observation. Those variables with a .1 (also denoted by the blue line and circles) represents the difference between male and female proportions using the respondent self report of their sex. All points include their respective 95th confidence interval around the male/female difference obtained in the test of proportions. Comparing the two points for each variable, all of the confidence internals overlap. This means that for all of the variables, there is no statistical significant difference between the estimates obtained using the interviewer observation compared with those obtained using the respondent report.

However, although final survey estimates may not be statistically different when using an interviewer observation, the graph does show that some substantive analyses may be affected. That is, researchers could draw different conclusions about whether or not a difference between males and females exists. Take, for example, attribute 18. This is the same variable discussed above (concern that family may need to turn to food assistance). The CI for attribute18.2, the difference in estimate between observed males and female, crosses the 0 line – insinuating that there is not a statistically significant gender difference in this variable. However, looking at the CI for attribute 18.1, a researcher would see that the 0 line is not included, thus, perhaps there is actually a difference between males and females when it comes to their report of needing to turn to food assistance. This situation, where the interviewer observed gender data may lead a researcher to draw a different conclusion, occurs in 7 of the 24 variables (29%).

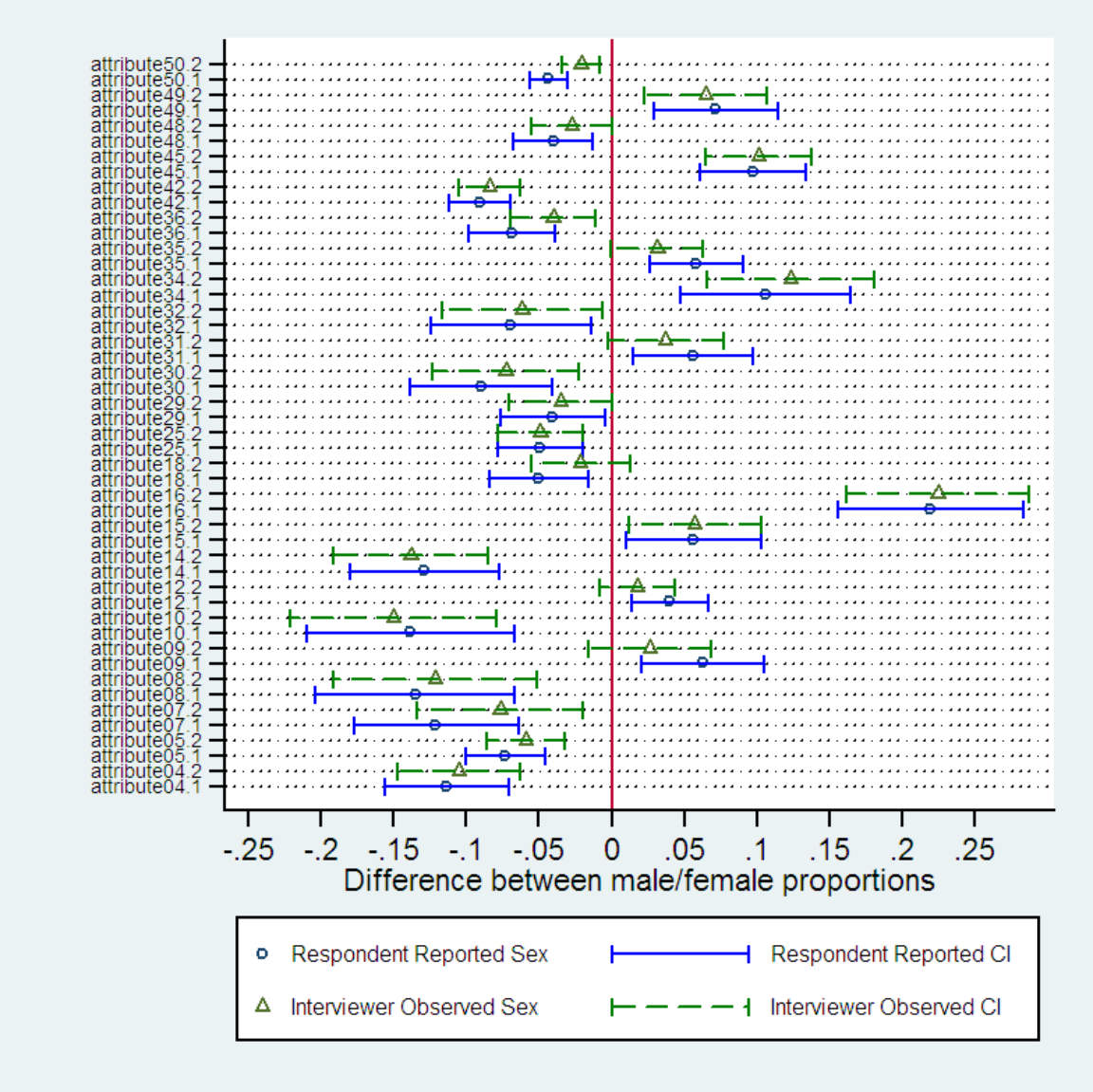


Figure 4.06. Difference Between Male and Female Proportions Using Both Gender Measures Where a Statistical Male/Female Difference Does Exist

Figure 4.07 presents the difference between male and female proportions using both gender measures for the 26 of the 50 variables that did not have a significant difference in self-reported male/female estimates. All of the CI's again overlap, showing there is not a significant difference in the estimates obtained using the gender observation

vs. the respondent report. As expected, these variables are affected less by the misclassification.

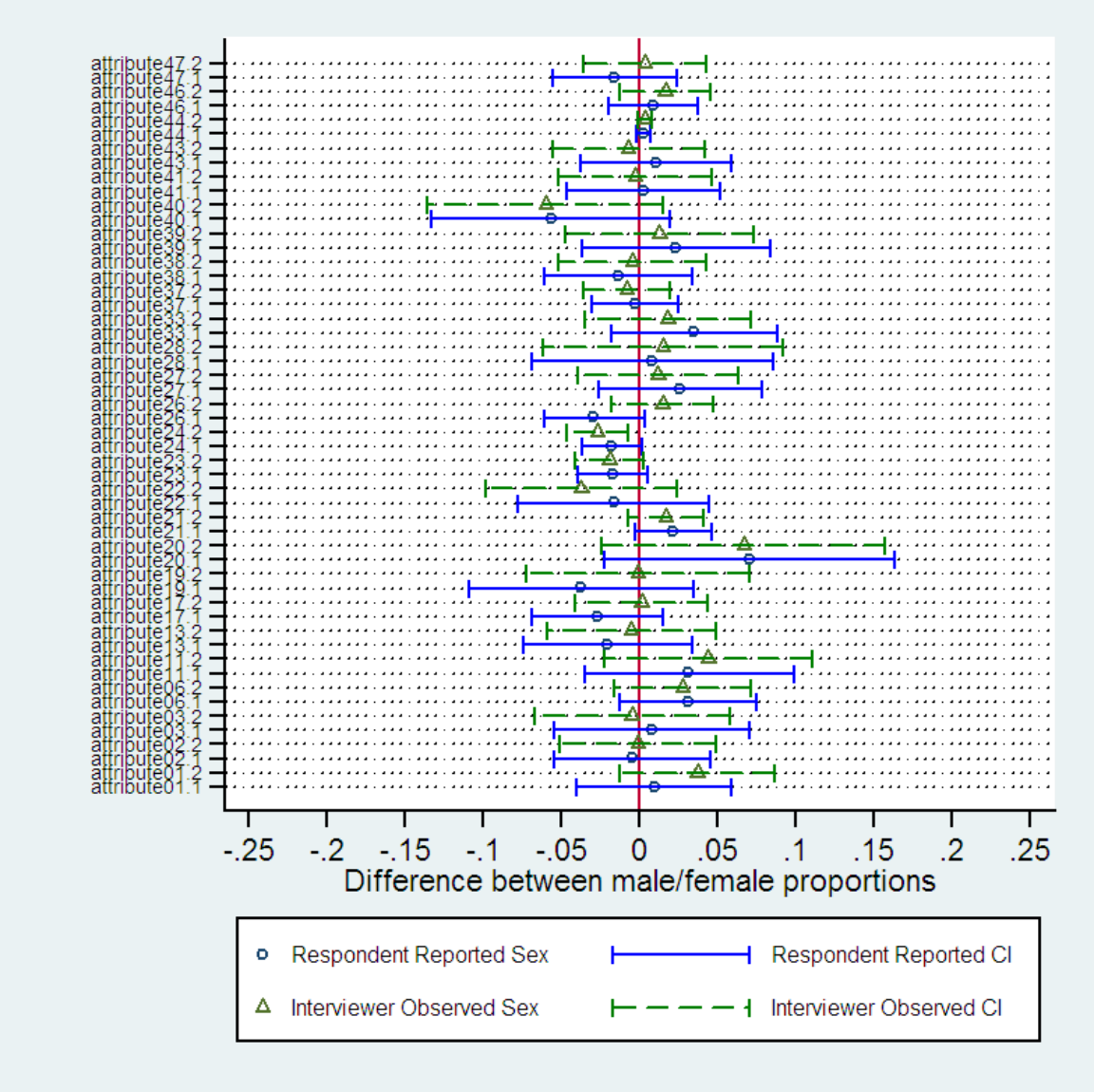


Figure 4.07. Difference Between Male and Female Proportions Using Both Gender Measures Where a Statistical Male/Female Difference Does Not Exist

4.4. Discussion

This chapter of my dissertation found small significant differences when comparing survey estimates by gender using the respondent's self-report (assumed to be the true-value) versus the interviewer observation of gender. The results do show that misclassification in interviewer gender observations can lead to some bias when gender differences exist, especially in male estimates since females are more likely to be misclassified. In terms of the size of shifts in estimates, the findings do provide some cautiously optimistic news for data collection firms that utilize observational methods to collect gender, especially if they have large sample sizes. However, the research should caution analysts that inferences drawn from differences in attitudes between males and females, made using gender judgments, could be affected by observational errors.

Despite finding no profound, statistical shifts in estimates, the research does call for evaluating the practical significance of the differences. Let's, for example, consider the concern for needing to turn to food assistance measure in the dataset. Using the interviewer observation of gender, 14% of adult males in the U.S. reported being very concerned; 12% show this level of concern when using the respondent self-report. While this may be only a two percentage point different, when extrapolating the numbers to the United States adult male population, this means that over 2 million males could be over or under counted depending on the gender collection approach. Research documenting incidence rates of certain behaviors diseases, or needs for males and females, such as the BRFSS, may find that even extremely small shifts in estimates leads to different conclusions and implications.

Given Chapter 3 of this dissertation found that increased exposure to respondent's voice increased the accuracy of gender observation, a limitation of this analysis is that it only included interviewer observations made very early in the survey (only available data). Future work could address this by using gender observations made early and late in the survey to identify if the increased accuracy of the observations eliminated measurement error in the final estimates. Another shortcoming of this work is related to the use of t-tests to identify significant differences between sample means. I recognize that this method of testing could lead to type 1 errors. A way to address this in the future might be to use the Bonferroni correction.

5. Discussion

Respondent gender is a widely used data point to understand differences in behaviors and attitudes. Most social science research includes gender in some form, either as a primary explanatory variable or as an independent, control variable. In addition, survey methodologists may use gender data to screen for survey eligibility, inform survey logic, contribute to nonresponse assessment and adjustments, and design experimental research. Reviews of survey documentation and anecdotal evidence suggested that many survey research organizations collect gender by some form of interviewer observation. Research in the paradata literature, as well as linguistics show, however, that such observations are imperfect. In fact, preliminary research in this dissertation suggested an overall misclassification rate of 8% in gender observations and higher for women and African-Americans.

Despite widespread uses and applications of gender data, no known research has explored any predictors of observational errors in interviewer judgments of respondent gender such as the length of exposure to the voice or the noisy environment of a typical centralized phone room. Moreover, we know nothing about the potential consequences to final survey estimates in collecting respondent gender by observation. Until this dissertation, these were significant gaps in the literature with both theoretical and practical implications for survey practitioners and methodologists. .

Using primary data collection, experimental methods, and secondary data analysis, this dissertation provided the first documentation of (1) how respondent gender is collected and used by firms; (2) how length of exposure and the presence of noise

contribute to observational errors; and (3) whether errors in gender observations affect final survey estimates.

5.1. Dissertation Findings

Chapter 2 discusses the findings from a survey of firms who perform telephone data collection. The goal was to identify the practices in the collection and uses of respondent gender information. Variations of collecting respondent gender include: (1) asking the respondent (e.g., *Are you male or female?*); (2) interviewer observation only (e.g., *Interviewer: Record the gender of the respondent*.); (3) a hybrid of both observation and respondent report, called *ask-assisted* throughout this dissertation, where firms instruct interviewers to ask respondents their gender only when necessary or not obvious (e.g., *Interviewer: Record the gender of the respondent. Ask only if not obvious*.); (4) utilizing record or administrative data; and (5) multiple methods (e.g., a combination of the above approaches). The key finding is that only 15% of firms ask respondents their gender directly. The majority of firms (68%) obtain gender through some form of interviewer observation – 30% doing so by a pure observation and 38% using a hybrid of both observation and respondent report, called *ask-assisted*, where firms instruct interviewers to ask respondents their gender only when necessary or not obvious (e.g., *Interviewer: Record the gender of the respondent. Ask only if not obvious*.). Among the firms that collect by observation (either purely or ask-assisted), 45% most often do so early in the telephone survey as part of the introduction or screening, leaving the interviewer little time to listen to vocal cues. This placement provided motivation for the subsequent experimental work. If I had found in this chapter that the collection of gender by observation was rare, the implications and practical relevance of my dissertation could

have been minimal. Finding many firms not only rely on interviewer observations for respondent gender collection but use them in a wide range of applications, these baseline measures were an important backdrop for the rest of the research.

Chapter 3 identified and experimentally tested two situational sources of error in interviewer observations of respondent gender: exposure length (Does how much time survey interviewers have to observe whether they are speaking with a male or a female affect the quality of their judgment?) and the presence of noise (Does the noisy environment of a centralized phone room contribute to observational errors?) Recordings from previously conducted survey interviews were stratified by males and females, white and non-white respondents, and hard and easy cases (determined by an incorrect/correct gender assignment by the initial interviewer). Twenty-seven raters were exposed to respondent's voice at various length (one second, five second, thirty second, full survey exposure) and two noise treatments (no auxiliary noise; phone room noise overlaid). Of all one second recordings, 28.50% were miscoded. A steady decline in respondent gender misclassification was observed as time exposed to the voice increased: 22.51% of recordings that were 5 seconds in length were assigned an incorrect gender; 17.82% of those with 30 seconds of respondent voice were misclassified. Gender misclassification did not completely disappear with full exposure to the speaker's voice. Even after hearing the entire survey recording, 12.27% of these cases were still incorrectly observed.

The noise treatment showed findings different from what was anticipated. The presence of noise actually improved the accuracy of gender observations: 17.61% of all cases with the noise overlay were miscoded; 22.91% of the cases without noise were miscoded. One can only speculate about the reasons for this result. However, as

discussed in greater detail at the end of Chapter 3, one explanation includes the increased focus and attention interviewers use when faced with a noisy signal.

In terms of the consequences of observational errors, analyzing secondary data, Chapter 4 found little differences when comparing survey estimates by gender using the respondent's self-report (assumed to be the true-value) versus the interviewer observation of gender. Differences between the proportions obtained with the two measures for the binary outcomes are small: an average absolute value of 2.9 percentage points for males; average absolute value of 1.9 percentage points for females. Overall, there is greater misclassification of females. As a result, male estimates are affected more than female estimates by interviewer observational errors in gender classification. This is because there are more males with female responses than true females with male responses when using observational data. Relative bias is also fairly small for attitudinal, behavioral, and demographic variables – average for males and females is 4.26% and -1.56%, respectively, in the attitudinal variables and slightly less for the behavioral and demographic outcomes. The findings provided some good news for data collection firms that utilize observational methods to collect gender. However, this good news is tempered by the fact that researchers could make different conclusions about male and female response patterns depending on how respondent gender was collected. In other words, it is possible that using an interviewer observation to test for male/female differences could lead to different conclusions than if the respondent report was used in testing.

Looking at the preliminary research and sets of analyses collectively, the findings show that telephone data collection organizations frequently use interviewer observations

to obtain respondent gender. Although the quality of these observations is imperfect and is affected by characteristics of the situation in which they are being collected, they are widespread used. The promising news is that imperfections cause little shifts in male and female survey estimates.

5.2. Contributions of this Dissertation

This dissertation has both practical and theoretical contributions. In terms of contributions to practice, this dissertation provides survey practitioners guidance in the use and collection of interviewer gender observations. This work provides (1) base rates of misclassification that can inform whether firms should continue to feel comfortable with documented errors; (2) information about how industry firms collect and apply gender data which can be used to brainstorm alternative approaches and applications; (3) rates of observational misclassification of gender at various lengths of exposure to the voice, which can be used to inform survey design decisions; (4) rates of misclassification in the presence or absence of noise, which firms can consider when making organizational decisions and data collection procedures; and (5) implications for male and female survey estimates for practitioners to evaluate whether the observational errors are consequential enough to move to respondent self-reports.

Theoretically, findings from this dissertation will contribute to several bodies of literature. In linguistics, this work enriches existing knowledge related to the quality of judgments using only acoustic cues. Given the experimental findings in Chapter 3, linguists can use this work to motivate the exploration of how vocal features (other than pitch) may contribute to gender misclassification errors. Furthermore, I document error rates in listener observations of speaker gender outside of a standard lab environment,

used by most of the linguists. This may strengthen the external validity of some of their findings. Also, this dissertation contributes to a greater understanding of how speaker and listener characteristics correlate with observations. Limited research had been conducted to understand how increased listening time improves the quality of judgments. This dissertation may provide a guide for linguists to explore longer exposure lengths.

For survey methodologists, this dissertation adds new information to the paradata literature showing the utility and quality of observational data in a telephone setting. Limited work has been conducted until now to understand how these paradata can be collected and used outside of face-to-face data collection modes. It provides a foundation for the exploration of other types of aural observational data such as race and age and the evaluation of whether it can serve as an effective nonresponse adjustment. This dissertation also contributes to the literature addressing interviewer data quality and their differential ability to collect paradata.

This dissertation may also contribute to the psychology literature showing the role that stereotyping has on judgments (e.g., low vocal pitch is assumed to be a male) and extend existing work related to vocal judgments. Through the discussion of speaker and listener characteristics, social psychologists can further explore contributors of stereotyping and thin-slice judgment formation.

5.3. Recommendations and Next Steps

As part of the discussion in each of the three studies presented, I addressed limitations and implications specific to that work. Here, I discuss more general recommendations and next steps of the dissertation research.

Survey methodologists routinely examine the impact of question design features and experiment with various constructions of questions, while recognizing the important role of demographic questions in research. As such, certain questions—such as those collecting income (e.g., Moore et al. (2000)) and race (e.g., Bayer (1973))—have been given a lot of attention in the literature and at conferences. No known work has been done on gender collection. Research has shown that the best way to avoid measurement error is to ask, rather than observe respondent behaviors or characteristics. But, as shown in this dissertation, the norm in practice is not to ask. Firms that observe gender feel it is uncomfortable and unnecessary to ask. Anecdotally, when talking with interviewers who are accustomed to only observing gender, the idea of changing to asking for gender is often met with strong reluctance and discomfort. Interviewers frequently use strategies to overcome the awkwardness (e.g., inserting hedges, qualifiers, and apologies before asking whether the respondent is a male or female), but little is known about how these strategies work. Experimentation on how to simply ask "are you a male or female" may be less beneficial from experimentation related to the use of qualifiers such as: "*I'm required to ask, are you male or female*?"; "*We need a verbal response for everything; are you a male or female?; "I apologize, I have to ask, are you a male or female?*" Do variations lead to different item nonresponse or break-offs? Experimental work may be able to find a solution to increasing the accuracy of the data from an observation while tempering perceived offensiveness of asking respondents their gender.

In Chapter 2, I found that among the primary reasons why firms do not ask respondents their gender is for efficiency. Many survey data collection firms mentioned that eliminating this item from the questionnaire saves valuable times and space in the

instrument, reducing length, respondent burden, and thus cost. However, what if observing gender actually increases the survey administration length, thus giving interviewers more time to observe gender? Interviewers need to listen to respondent's voices, disentangle confusing signals, make their judgment, record their judgment, and then move to the next question. Is the amount of time needed to perform these functions really saving time? Future work might evaluate the time it takes for interviewers to simply ask gender compared with observing. A cost-savings may not exist. Furthermore, assuming interviewers are silent through their observation process, do these momentary pauses induce any break-offs or inhibit the important process of rapport building early in the survey? It might be less disruptive for firms to ask.

Singer, et al. (1983) suggests that interviewer bias is a significantly greater threat in telephone surveys, because usually fewer interviewers complete more surveys and workload is higher when compared with other data collection modes. The preliminary work discussed in Chapter 1 did find some evidence for the clustering of errors with certain interviewers. However, the experimental research in Chapter 3 did not come to the same conclusion. Similar findings were obtained by Oksenberg et al. (1986) which found rater agreement was typically high and reflects vocal stereotypes, not necessarily accurate judgments. It is likely that the experimental setting affected this outcome (e.g., use of more experienced interviewers, increased focus and attention to detail in the lab setting), however, more work is needed to disentangle whether certain interviewers are more adept at making observations. If observational errors do cluster around specific interviewers, additional training may be able to decrease the error rate. Also, practitioners

should consider these issues and limit the number of completed cases for each interviewer.

As mentioned in Chapter 1, Smith (1997) conducted research to evaluate the quality of interviewer observations of respondent race in a face-to-face setting (the General Social Survey). Although the racial misclassification was fairly small (at most 5.5%), Smith concluded that changing the GSS race measurement to self-identification would make it more in line with practices used in other major surveys. However, he also noted that this change in methodology could disrupt longitudinal comparisons. Researchers have long known that changes in ways of asking or collecting the same measure, could lead to different estimates. This could be a consideration for researchers when evaluating their gender data collection practices. For example, in the large-scale national telephone survey BRFSS, the ask-assisted gender collection method was used. Moving to a respondent self-identification methodology might show slight differences in gender analyses over time. While this is unlikely to be problematic for gender, firms should consider it when conducting a longitudinal study or simply provide documentation related to the methodological change in their approach.

Smith's (1997) discussion about the trade-offs of collecting respondent race by observation versus self-identification has similar applications to gender. He suggests that while self-identification may increase item non-response (e.g., higher refusals) or lead to responses that do not fit into one of the researcher's categories, interviewer observations are not without error. Moreover, observations can only be used when there are "gross distinctions between physically distinct groups" (p. 5). This means that observations should not replace respondent self-reports if there are other, perhaps cultural,

considerations when defining the categories. This rationale is relevant to my dissertation when distinguishing respondent sex and respondent gender. As Walker and Cook (1998) state, "sex refers to the anatomical or chromosomal categories of male and female. Gender refers to socially constructed roles that are related to sex distinctions" (p. 255). I use these terms interchangeably since it is unclear whether asking or observing if a respondent is a male or female is looking to obtain gender or sex. However, I recognize that these terms take on different meanings and could have measurement implications. Using Smith's argument here, while *sex* creates distinct groups of males and females, *gender* distinctions may be less clear. How does an interviewer parse out possibly confusing gender signals? For example, a transgender respondent may differentiate between their self-identified gender and their biological sex. Perhaps survey researchers need to start considering whether they are looking to obtain a respondent's gender or sex. How does this decision then affect measurement error in interviewer observations of their sex/gender compared with self-reports?

Along these lines is the need to evaluate this issue with a cross-cultural/multi-national sensitivity. I address only practices in the collection of respondent gender with firms located in the United States. In some cultures, asking respondents their gender may be offensive whereas others may not find it at all troublesome. Focus groups or field tests in various cultures and countries could inform best practices.

The preliminary work discussed in Chapter 1, recordings used in Chapter 3, and data analyzed in Chapter 4, used observations obtained from landline surveys. While Chapter 3 found that noisy signals did not induce higher misclassification, gender misclassification by observation should be tested in cell phone surveys. Poor call quality

and the environment in which a respondent is taking the survey may affect observational error rates in cell phone cases.

While Chapter 4 found a minimal impact on final survey estimates, there are other possible consequences which should be explored in future research. Survey break-offs and item non-response are a likely result if flawed observations are used to determine survey logic. For example, if a respondent is observed to be a male when her gender is a female, she may be given questions related to traditional male topics (e.g., prostate cancer in the BRFSS). How respondents handle these situations is unclear. Does she simply end the call? Does she say that she is female and if so, does the interviewer go back and change his or her observation? Do interviewers just skip through the inappropriate questions, never changing their initial observation, thus making analysis difficult since gender information may be inconsistent? Additional work is needed to address the implications and possible additional consequences of observational errors.

As mentioned in Chapter 1, the paradata literature is a fairly recent contribution. Much has yet to be explored, especially the quality of observational data and non-face-to-face modes. This dissertation focused on only one form of observational paradata, namely gender data. Future research could extend this work to other forms of observational data. What is the efficiency of observing other demographic measures in telephone surveys?

Emerging work within the paradata literature is exploring how respondent and interviewer vocal properties affect survey participation and outcomes (e.g., Conrad et al., 2013). However, much work is needed when it comes to evaluating the vocal properties that inform telephone interviewer judgments. While Chapter 3 of this dissertation was

one of the first attempts at addressing this issue (looking only at average pitch), there is a realm of possibilities for future research. The linguistics literature points out that females tend to have more variable speech (using a wider pitch range than males). It would be interesting to explore this pitch variability and understand if gender assessments are a result of the "average" signals or taken from the variability (vocal extremes). Could we predict misclassification from other measured respondent vocal properties (e.g., jitter, breathiness, energy and power of the voice)?

Throughout this dissertation, the respondent self-report is assumed to be a true-value. However, respondents could choose to give a wrong answer in an effort to protect their anonymity or confidentiality. The validity of this assumption could be evaluated in future research. An experimental approach (with unlimited time, money, and access to records) to test whether respondents ever misrepresent their gender might be to compare observations with administrative records. Ideally, I would obtain administrative records from a source that is likely to include people with a range of age, race, ethnicity, and gender (e.g., not records from a college). Females, older people, or certain cultures, for example, might be more likely to not reveal their true gender. The administrative records might be from a doctor's office where gender is biologically confirmed data. I would not want to use administrative records that are simply aided by a name since names are often androgynous. Records would be randomly selected and telephone interviewers would contact selected respondents to participate in a survey, with the primary goal of collecting an interviewer observation and respondent self-report of gender. Respondents would not be made aware of the actual goal of the study since it would likely alter their behavior. One-third of the observations and respondent reports would be done early in the survey,

one-third in the middle of the survey, and the remaining one-third done at the end of the survey. Like Chapter 3, this would allow for the evaluation of whether time increases the accuracy of self-reports (due to, perhaps, increased rapport building). In addition, it could reveal whether asking for such demographic information early in the survey causes break-offs. For this study, I would not alter the wording of how to ask for the self-report (e.g., Are you male or female?).

I conclude this work with the recommendation that survey firms should ask, not observe, respondent gender. The documented misclassification rates are notable and there are some clear predictors of inaccurate observations. Firms that need gender information early in the survey to inform survey logic should be especially open to using self-reports rather than observational data. Although final survey estimates comparing male and female outcomes may not be greatly affected by the use of an interviewer observation or self-report, different conclusions could be made. Like many design decisions, survey researchers will have to weigh the pros and cons of asking. It may be slightly uncomfortable to ask for gender when it seems obvious, but why take a chance that an error could result in measurement error or final biased estimates? Organizations that have traditionally collected gender by observation may meet this change with some resistance by interviewers. Perhaps interviewers will feel it is unnecessary and uncomfortable. If so, proper training to address the concerns is needed. Additionally, adequate supervision and monitoring throughout the transition is necessary to provide feedback and to ensure interviewers are actually asking the question. Training protocols would be greatly enhanced by the use of recordings, which present both hard and easy

male and female voices as was done in Chapter 3. Interviewers may benefit from hearing characteristic of voices that are commonly assigned an incorrect gender.

Appendix A: Questionnaire for Survey of Data Collection Firms

Q1. Does your organization conduct telephone surveys?*

○ 1. Yes

○ 2. No *<Terminate>*

Q2. In telephone surveys, which of the following best describes how your organization <u>most often</u> collects the gender of the respondent?*

○ 1. Interviewer observation only: interviewer never asks gender (e.g., "Interviewer: Record gender of respondent")

○ 2. Interviewer observation but interviewer asks the respondent when necessary or not obvious (e.g., "Interviewer: Record gender of respondent. Ask 'Are you male or female' if not obvious.")

○ 3. By always asking the respondent (e.g., "Are you male or female?")

○ 4. By multiple methods or a combination of the above methods

○ 5. By some other method (please describe) [ ]

○ 6. Do not collect respondent gender *<Terminate>*

<If Q2=4>

Q3. You indicated that your organization <u>most often</u> collects the gender of the respondent by multiple or a combination of methods. Please select all of the methods that apply.

☐ 1. Interviewer observation only: interviewer never asks gender (e.g., "Interviewer: Record gender of respondent")

☐ 2. Interviewer observation but interviewer asks the respondent when necessary or not obvious (e.g., "Interviewer: Record gender of respondent. Ask 'Are you male or female' if not obvious.")

☐ 3. By always asking the respondent (e.g., "Are you male or female?")

☐ 4. By some other method (please describe) [ ]

Q4. Does your organization always collect gender using this method or does the method vary depending on the project or client?

○ 1. Always the same way

○ 2. Varies across projects or clients

Q5. Does your organization always collect gender in the same place in the questionnaire or does the location vary depending on the project or client?

○ 1. Always in the same place

○ 2. Placement varies across projects or client

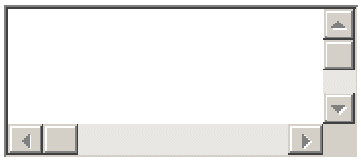
<If Q2=1,2,4>

Q6. In your telephone surveys, which of the following best describes where in the survey interviewers <u>most often</u> observe the respondents' gender?

- 1. In the survey introduction or screening
- 2. In the middle of the survey
- 3. At the end of the survey

<If Q2=1,2,4>

Q7. What instructions, if any, are interviewers given in the questionnaire when <u>observing</u> the gender of the respondent? Please be specific and, if possible, copy and paste the actual instructions.



<If Q2=1,2,4>

Q8. Please indicate how often, if ever, your organization typically uses interviewer observations of a respondent's gender in each of the following ways.

	For all studies 1	For many studies 2	For few studies 3	Never 4
a. To assign skip patterns/inform survey logic	○	○	○	○
b. For weighting purposes	○	○	○	○
c. Other forms of non-response adjustment	○	○	○	○
d. For interviewer tailoring or accommodation strategies	○	○	○	○
e. Screening to determine eligibility for participation in the survey	○	○	○	○
f. As a substantive variable used in	○	○	○	○

analyses and reports based on the survey

<If Q2=1,2,4>

Q9. Are there any other ways that your organization ever uses interviewer observations of a respondent's gender?

- 1. Yes (please describe) _____
- 2. No

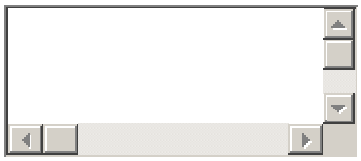
<If Q2=3,5>

Q10. Please indicate how often, if ever, your organization typically uses gender data collected in a survey in each of the following ways.

	For all studies 1	For many studies 2	For few studies 3	Never 4
a. To assign skip patterns/inform survey logic	○	○	○	○
b. For weighting purposes	○	○	○	○
c. Other forms of non-response adjustment	○	○	○	○
d. For interviewer tailoring or accommodation strategies	○	○	○	○
e. Screening to determine eligibility for participation in the survey	○	○	○	○
f. As a substantive variable used in analyses and reports based on the survey	○	○	○	○

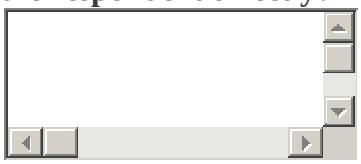
<If Q2=1>

Q11. Please describe the reason(s) why your organization typically collects gender by interviewer observation only instead of asking the respondent directly? Please be as specific as possible.



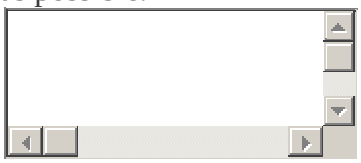
<If Q2=2>

Q12. Please describe the reason(s) why your organization typically collects gender by asking the respondent only when interviewers feel it is necessary instead of always asking the respondent directly? Please be as specific as possible.



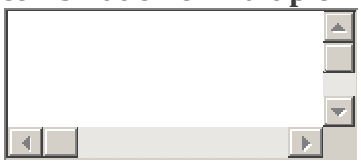
<If Q2=3>

Q13. Please describe the reason(s) why your organization typically collects gender by directly asking the respondent instead of by interviewer observation? Please be as specific as possible.



<If Q2=4>

Q14. Please describe the reason(s) why your organization collects gender <u>both</u> by a combination or multiple methods? Please be as specific as possible.



<If Q2=1,2,4>

Q15. Does your organization provide specific interviewer training on the <u>observation</u> of respondents' gender?

- 1. Yes
- 2. No

<If Q15=1>

Q16. Please describe the interviewer training that your organization provides to interviewers on the observation of respondents' gender.

<If Q2=2>

Q17. You indicated that your organization most often collects gender data by asking the respondent only when interviewers feel it is necessary to ask. **What percent of the time do you think interviewers do, in fact, ask the respondent their gender?**

0 Enter 0-100 for percent

Q18. How accurate would you say interviewer observations of respondent gender are over the phone?

0 Enter 0-100 for percent accuracy

Q19. Does your organization routinely collect interviewer key stroke data in telephone surveys?

- 1. Yes
- 2. No
- 3. Unsure

Q20. Please indicate which, if any, other forms of data collection your organization conducts.

- 1. Face-to-Face surveys
- 2. Mail surveys
- 3. Web surveys

<If Q20=1>

Q21. You indicated that your organization conducts face-to-face surveys. **Please list what interviewer observations of household or respondent-level features, if any, your organization typically collects in face-to-face surveys. Please include a description of the purpose for collecting each observation.**

Q22. In which year was your organization founded?*

0 Year (YYYY)

Q23. What proportion of all of your organization's work was sponsored by:

Political clients

0 %

Media clients

0 %

Other commercial clients

0 %

The Federal government

0 %

Other government clients

0 %

Other not for profit clients

0 %

Academic clients

0 %

Total

%

Q24. Is your organization:*

- Commercial, non-marketing
- Commercial, marketing
- Academic
- Other, not for profit

Q25. Does your organization subscribe to the AAPOR code of ethics?

- Yes
- No

Q26. What is your job title?

Appendix B: Survey Invitation Text

Dear Colleague,

My name is Susan McCulloch and I am a PhD candidate in the Joint Program in Survey Methodology (JPSM) at the University of Maryland. I am hoping that you or someone in your organization can answer a very brief survey about your organization's data collection methods.

Follow this link to the Survey:
<survey link>
Or copy and paste the URL below into your internet browser:
<survey link>
If you think there is someone else in your organization that is better suited to complete the survey, I would greatly appreciate you forwarding this email to him or her.

The survey is integral to my dissertation research and will take less than 10 minutes of your time and although participation is voluntary, we all know how important those response rates can be! All data will be kept strictly confidential and no responses will ever be linked to specific respondents or organizations. Data will only be analyzed and reported in the aggregate.

Please feel free to contact me at skenney@survey.umd.edu or my dissertation chair, Dr. Frauke Kreuter, at fkreuter@survey.umd.edu with any questions or concerns. Also, please email me if you are interested in receiving a copy of the survey findings.

Many thanks for your time and help in the advancement of my research.

Best,
Susan

Susan Kenney McCulloch
PhD Candidate
Joint Program in Survey Methodology/University of Maryland

Appendix C: Description of Selected Variables for Analysis

Question	Full question wording	Sample Size	1 Value
attribute1	In general, do you think the city's bus and subway system is getting better, getting worse, or staying about the same?	775	Thinks NYC bus and subway system is getting better
attribute2	Would you rate the JOB Mayor Michael Bloomberg is DOING IN OFFICE as excellent, good, fair, or poor?	800	Rates Mayor Bloomberg's job in office as excellent
attribute3	From what you have read or heard, do you approve or disapprove of Congress providing federal loans to American automakers?	958	Approves of Congress giving federal loans to US automakers
attribute4	Should President Obama: (1) Address the health care crisis even if it means more government debt or (2) Not address the health care crisis?	1936	Prefers next president deals with health care crisis over cutting taxes
attribute5	Would you rate the job Senator Charles Schumer is doing in office as excellent, good, fair, or poor?	5134	Rates Senator Schumer's approval rating as excellent or good
attribute6	Should President Obama: (1) Cut taxes even if it means more government debt; (2) Cut taxes?	1935	Says Obama should cut taxes even if it means more debt
attribute7	Do you think a law that requires restaurants to show calorie counts for items on their menu should show nutritional information on a per serving basis or do you think restaurants should show calorie counts for the entire item, regardless of the number of servings?	1097	Thinks restaurants should be required to show calorie counts per serving
attribute8	Would you rate the job Senator Hillary Clinton has done in office as excellent, good, fair, or poor?	641	Rates Hillary Clintons job in office as excellent
attribute9	Do you approve or disapprove of how Mayor Michael Bloomberg is handling the city's budget?	2168	Approves of how Mayor Bloomberg is handling the city's budget
attribute10	Do you favor or oppose legalizing same-sex marriage in New York State?	739	Favors legalizing same-sex marriage in NYS

attribute11	Which one of the following issues do you think Michael Bloomberg has handled the BEST as mayor?	701	Says Blomberg has handled education best as mayor
attribute12	In general, thinking about the way things are going in New York City, do you feel things are going in the right direction or that things are going in the wrong direction?	5,642	Says things in NYC are going in the right direction
attribute13	Do you approve or disapprove of the job the Republicans in Congress are doing?	968	Approves of the job Republicans in Congress are doing
attribute14	What age do you think is a good age to start talking to children about money?	1,059	Thinks 5 and under is age to start talking to kids about money
attribute15	Right now, do you think the U.S. economy is getting better, getting worse, or staying the same?	1,070	Thinks US economy is getting better
attribute16	Do you think the number of U.S. troops in Afghanistan should be increased, decreased, or remain the same?	912	Thinks the number of US troops in Afghanistan should be increased
attribute17	Do you think it is very likely, likely, not very likely, or not likely at all that each of the following will happen on Election Day: That many people will be confused about how to cast their ballots	686	Thinks its very likely that people will be confused how to cast their ballots
attribute18	How concerned are you that you and your family may need to turn to food assistance?	1,713	Very concerned that family may need to turn to food assistance
attribute19	How often do you worry that your total family income will not be enough to meet your family's expenses and bills: always, sometimes, seldom, or never?	635	Always worries that family income will not be enough
attribute20	In general, do you have a favorable or an unfavorable impression of Harold Ford?	307	Has a favorable impression of Harold Ford
attribute21	Which stage do you think is most difficult time to raise a child: (1) Infancy until they can walk, (2) Toddlers or preschool age, (3) School age from 5 to 9, (4) Pre-teen 10 to 12, (5) Teenagers 13 to 19, (6) Adulthood 20 and older?	1,121	Thinks the hardest time to raise a child is from infancy through walking

attribute22	Overall, has Barack Obama met your expectations, exceeded your expectations or fallen below your expectations as president?	1,029	Says Obama has met their expectations
attribute23	Do you think the current economic conditions are mostly something President Obama inherited or are they mostly a result of his own policies?	6,366	Thinks economic conditions were inherited by Obama
attribute24	In general, thinking about the way things are going in the country, do you feel things are going in the right direction or that things are going in the wrong direction?	9,523	Says that things in the country are going in the right direction
attribute25	Are you very concerned, concerned, not very concerned, or not concerned at all that you or someone in your household will get the H1N1virus known as the swine flu?	1,758	Are very concerned about getting H1N1 flu virus
attribute26	Do you think the schools in your community are very prepared, prepared, not very prepared or not prepared at all to deal with a major disaster?	850	Thinks schools in community are very prepared to deal with a disaster
attribute27	Have you ever had a vacation that you considered to be a disaster?	850	Says they have had a vacation considered to be a disaster
attribute28	Do you think Tiger Woods should deal with the recent events that have been in the news only privately or do you think he should address them publically?	568	Says Tiger Woods should address the recent events in the news privately
attribute29	Do you think there should be a law to ban texting while driving?	923	Thinks there should be a law banning texting while driving
attribute30	Would you rate how Michelle Obama is doing as first lady as excellent, good, fair, or poor?	1,091	Rates Michele Obama's role as excellent
attribute31	Have you done any of the following things recently to reduce your own spending or save money?	1,116	Has done something recently to reduce spending money

attribute32	Every ten years the United States Census Bureau conducts a census of all Americans. 2010 is a census year. How likely are you to fill out the 2010 census form? Are you: very likely, likely, not very likely, not likely at all?	1,008	Says they are very likely to complete the 2010 Census form
attribute33	How many times in an average week do you eat dinner at home?	579	Eats dinner at home four or fewer days a week
attribute34	Do you plan on watching all, most, some, or none of the Super Bowl this coming Sunday?	1,019	Plans on watching all of the Super Bowl
attribute35	If November's presidential election were held today, whom would you support if the candidates are:	3,706	Plans to vote for McCain-Palin in 2008 election
attribute36	If the 2010 election for GOVERNOR of New York State were held today, whom would you support if the candidates are:	4,410	Will vote for Cuomo in the 2010 NYS Gov election
attribute37	If the 2010 election for GOVERNOR of New York State were held today, whom would you support if the candidates are:	4,966	Plans to vote for Paterson over Lazio in 2010 Governor race
attribute38	This month, the Winter Olympics will be in Vancouver, British Columbia Canada. Do you plan to follow the events: a great deal, a good amount, a little, or not at all?	1,023	Plans to watch a great deal of the winter Olympics in Vancouver
attribute39	What will you mostly do with your refund: Will you:…?	567	Plans to spend tax refund rather than pay bills or save
attribute40	Do you plan on watching or listening to a great deal, some, not too much, or none of the vice presidential debate next Thursday night between Sarah Palin and Joe Biden?	680	Plans to watch a great deal of Palin-Biden debate
attribute41	Approximately how often did you go to any live performances, such as plays, concerts, musicals, or dance, during the past 12 months? Would you say very often, fairly often, not very often, or not at all?	1,104	Went to live performances such as plays often
attribute42	Are you registered to vote as:	7,962	Reports being a Democrat

attribute43	Do you consider yourself and environmentalist or not?	1,098	Considers themselves an environmentalist
attribute44	Are you of Hispanic or Latino origin?	22,885	Is a Latino
attribute45	Are you currently employed, looking for work, not employed, or retired?	2,860	Is employed
attribute46	Do you, personally, have a Twitter account?	1,100	Has a Twitter account
attribute47	Do all, some, or none of the adults in your household have health insurance or a health plan right now?	1,166	Is a household with health insurance
attribute48	Is your combined family income before taxes:		Has an annual income of less than $15000
attribute49	In which year were you born (recoded into years)?	1,744	Is under 45 years old
attribute50	Are you Protestant, Catholic, Jewish, or Muslim?	22,217	Is a Protestant

References

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

Ambady, N., & 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., Conner, B. (1999). Accuracy of judgments of sexual orientation from thin slices of behaviors. *Journal of Personality and Social Psychology*, *77*(3), 538-547.

Ambady, N., & Hallahan, M., Rosenthal, R. (1995). On judging and being judges accurately in zero-acquaintance situations. *Journal of Personality and Social Psychology*, 69(3), 518-529.

Andrews, M., & Schmidt, C. (1997). Gender presentation: perceptual and acoustical analyses of voice. *Journal of Voice, 3*, 307-313.

Bailey, L., Moore, T., & Bailar, B. (1978). An interviewer variance study for the eight impact cities of the National Crime Survey cities. *Journal of the American Statistical Association, 73*(361), 16-23.

Bagnall, A.D., Dorrian, J., & Fletcher, A. (2011). Some vocal consequences of sleep deprivation and the possibility of "fatigue proofing" the voice with voicecraft voice training. *Journal of Voice, 25*(4), 447-61.

Bayer, A. (1973). Construction of a Race Item for Survey Research. *The Public Opinion Quarterly*, 36(4), 592-602.

Boone, D. (1997). Is your voice telling on you?: how to find and use your natural voice. San Diego: Singular Publishing Group.

Bose, J. (2001). Nonresponse bias analyses at the National Center for Education Statistics. *Proceedings of the Statistics Canada Symposium*. Retrieved September 27, 2012, from http://www.fcsm.gov/committees/ihsng/StatsCan2_JB.pdf.

Brend, R M. (1975). Male-female intonation patterns in American English. In B. Thorne & N. Henley (Eds.), *Language and sex: differences and dominance* (pp. 84-87). Rowley: Newbury House Publishers.

Bull, R., Clifford, B.R. (1984). Earwitness voice recognition accuracy. In G.L. Wells & E.F. Loftus (Eds.), *Eyewitness Testimony: Psychological Perspectives* (pp. 92-123). New York: Cambridge University Press.

Callegaro, M., De Keulenaer, F., Krosnick, J., & Daves, P. (2005). Interviewer effects in a RDD telephone pre-election poll in Minneapolis 2001: An analysis of the effects of interviewer race and gender. *Proceedings of the Survey Research Methods Section of the American Statistical Association, 3815*-3821.

Casas-Cordero, C. (2010). *Neighborhood characteristics and participation in household surveys*. (Doctoral dissertation). University of Maryland.

Coleman, R. (1976). A comparison of the contributions of two voice quality Characteristics to the perception of maleness and femaleness in the voice. *Journal of Speech & Hearing Research*, *19*, 168-180.

Conrad, F., Broome, J., Benki, J., Kreuter, F., Groves, R., Vannette, D., McClain, C. (2013 forthcoming). Interviewer speech and the success of survey invitation. *Journal of the Royal Statistical Society*, *176(1)*, 1-20.

Copas, A., & Farewell, V. (1998). Dealing with non-ignorable nonresponse by using an 'enthusiasm-to-respond' variable. *Journal of the Royal Statistical Society – Series A, 161*(3), 385-396.

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.

Damste, P.H. (1964). Virilization of the voice due to anabolic steroids. *Folia Phoniatrica*, *16*, 10-18.

Davern, M., Rodin, H., Beebe, T., & Call, K. (2005). The effect of income question design in health surveys on family income, poverty and eligibility estimates. *Health Services Research*, *40*(5), 1-19.

Eckman, S. (2010). *Errors in housing unit frames and their effects on survey estimates*. (Doctoral dissertation). University of Maryland.

Eckman, S., & Kreuter, F. (2012). Undercoverage rates and undercoverage bias in traditional housing unit listing. Unpublished manuscript.

Goggin, J. P., Thompson, C. P., Strube, G., & Simental, L. R. (1991). The role of language familiarity in voice identification. *Memory and Cognition*, 19, 448-458.

Goldstein, U. (1980). *An articulatory model for the vocal tracts of growing children*. (Doctoral dissertation). Massachusetts Institute of Technology.

Groves, R.M., & Heeringa, S.G. (2006). Responsive design for household surveys: Tools for actively controlling survey errors and costs. *J.R. Statist. Soc. A, 169,*

Part 3, 439-457.

Groves, R.M., & McGonagle, K.A. (2001). A theory-guided interviewer training protocol regarding survey participation. *Journal of Official Statistics*, *17*(2), 615-626.

Groves, R.M., Wagner, J., & 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 of the American Statistical Association*, 3428-3431.

Graddol, D., & Swann, J. (1989). *Gender voices*. Cambridge: Basil Blackwell.

Hawkins, F. (1993). *Speaker ethnic identification; the roles of speech sample, fundamental frequency, speakers and listener variations.* (Doctoral dissertation). University of Maryland.

Haas, A. (1979). Male and female spoken language differences: Stereotypes and evidence. *Psychological Bulletin, 86*(3), 615-626.

Hagood, M. & Price, D. (1952). Statistics for sociologists. New York: Holt, Rinehart and Winston.

Hanson, H. (1997). Glottal characteristics of female speakers: Acoustic correlates. *Journal of the Acoustical Society of America*, *101*(1), 466–481.

Harb. H., & Chen, L. (2005). Voice-based gender identification in multimedia applications. *Journal of Intelligent Information Systems*, *24*(2-3), 179-198.

Harms, L. S. (1961). Listener's judgments of status cues in speech. *Quarterly Journal of Speech*, *47*, 164-68.

Hess W. (1983). *Pitch determination of speech signals*. New York: Springer-Verlag.

Hollien, H., Majewski, W., & Doherty, E. T. (1982). Perceptual identification of voices under normal, stress and disguise speaking conditions. *Journal of Phonetics*, *10*, 139-148.

Honorof, D., & Whalen, D. (2010). Identification of speaker sex from one vowel across a range of fundamental frequencies. *Journal of the Acoustical Society of America*, *128*(5), 3095–3104.

Hughes, S., & Rhodes, B. (2010). Making age assessments based on voice: The impact of the reproductive viability of the speaker. *Journal of Social, Evolutionary, and Cultural Psychology, 4*(4), 290-304.

Hyman, H., Cobb, W., Feldman, J., & Hart, C. (1954). *Interviewing in Social Research*. Chicago, IL: University of Chicago Press.

Imhof, M. (2010). Listening to voices and judging people. *International Journal of Listening*, *24*, 19-33.

Jans, M. (2010). *Verbal paradata and survey error: respondent speech, voice, and question-answering behavior can predict income nonresponse*. (Doctoral dissertation). University of Michigan.

Johnson, K., Strand, E., & D'Imperio, M. (1999). Auditory visual integration of talker gender in vowel perception. *Journal of Phonetics, 27*, 359-384.

Klatt, D. H., & Klatt, L. C. (1990). Analysis, synthesis, and perception of voice quality variations among female and male talkers. *Journal of the Acoustical Society of America*, *87*, 820–857.

Krauss, R., Freyberg, R., & Morsella, E. (2002). Inferring speakers' physical attributes from their voices. *Journal of Experimental Social Psychology*, *38*, 618-625.

Kreuter, F., & Casas-Cordero, C. (2010). Paradata. G*erman Council for Social and Economic Data,* Working Paper Series, No. 136.

Kreuter, F., Olson, K., Wagner, J., Yan, T., Ezzati-Rice, T.M., Casas-Cordero, C., Lemay, M., Peytchev, A., Groves, R.M., & 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.

Lakoff, R. (1975). *Language and women's place*. New York: Harper Colophon Books.

Lass, N. J., Hughes, K. R., Bowyer, M. D., Waters, L. T., & Bourne, V. T. (1976). Speaker sex identification from voiced, whispered, and filtered isolated vowels. *Journal of the Acoustical Society of America, 59*, 675-678.

Lass, N.J., Mertz, P. J., & Kimmel, K.L. (1978). The effect of temporal speech alterations on speaker race and sex identifications. *Language and Speech, 21*, 279-290.

Lass, N.J., Tecca, J.E., Mancuso, R.A., & Black, W.I. (1979). The effect of phonetic complexity on speaker race and sex identifications. *Journal of Phonetics, 7*(2), 108-118.

Laver J, Trudgill P. (1979). Phonetic and linguistic markers in speech. In Scherer K.R., & Giles H. (Eds.), *Social markers in speech* (pp. 1-32). Cambridge, England: Cambridge University Press.

Lessler, J., & Kalsbeek, W. (1992). *Nonsampling error in surveys*. New York City: John Wiley and Sons.

Lepkowski, J., Axinn W., Kirgis N., West B.T., Kruger S.N., Mosher W., & Groves, R. (2010). Use of paradata in a responsive design framework to manage a field data collection. *NSFG Survey Methodology*. Working Papers. Report 10-012.

Loebach, J., Pisoni, D., & Svirsky, M. (2009). Transfer of auditory perceptual learning with spectrally reduced speech to speech and nonspeech tasks: Implications for cochlear implants. *Ear Hear, 30*(6), 662-674.

Loveday, L. (1981). Pitch, politeness and sexual role: an exploratory investigation into the pitch correlates of English and Japanese politeness formulae. *Language & Speech, 24*(1), 71-89.

Lynn, P. (2003). PEDAKSI: Methodology for collecting data about survey non-respondents. *Quality and Quantity, 37*, 239-261.

Lynn, P., Sala, E. (2004). The contact and response process in business surveys: lessons from a multimode survey of employers in the UK. *ISER Working Papers.* 2004-12.

Maitland, A., Casas-Cordero, C., & Kreuter, F. (2009). An evaluation of nonresponse bias using paradata from a health survey. *Proceedings of the Survey Research Methods Section of the American Statistical Association,* 370-378.

Majewski, W., Hollien, H., & Zalewski, J. (1972). Speaking fundamental frequency of polish adult males. *Phonetica, 25*, 119-25.

McCulloch, S., Kreuter, F., & Calvano, S. (2010, May 14). *Interviewer observed vs. reported respondent gender: Implications on measurement error*. Paper presented at the 2010 American Association for Public Opinion Research, Chicago, IL.

McCulloch (Kenney), S., & Presser, S. (2006, May 20). *Survey research ethics: practicing what we preach?* Poster presented at the 2006 American Association for Public Opinion Research, Montreal, Canada.

McQueen, A., Vernon, S., Meissner, H., Klabunde, C., & Rakowski, W. (2006). Are there gender differences in colorectal cancer test use prevalence and correlates? *Cancer Epidemiology, Biomarkers Prevention, 15*, 782-791.

Mendoza-Denton, N., & Strand, E. A. (1998). *Sociobiological ideologies of gender difference in phonetic research*. Unpublished manuscript. The Ohio State University.

Mood, C. (2010). Logistic regression: Why we cannot do what we think we can do, and what we can do about it. *European Sociological Review, 26*(1), 67–82.

Moore, J., Stinson, L., Welniak, E. (2000). Income measurement error in surveys: A review. *Journal of Official Statistics, 16*(4), 331-362.

Newman, M., Groom, C., Handelman, L., & Pennebaker, J. (2008). Gender differences in language use: An analysis of 14,000 text samples. *Discourse Processes, 45*, 211–236.

Nygaard, L. C., & Queen, J. S. (2000, May). *The role of sentential prosody in learning voices*. Paper presented at the meeting of the Acoustical Society of America, Atlanta, GA.

Oates, J.M., & Dacakis G. (1983). Speech pathology considerations in the management of transsexualism: A review. *International Journal of Language & Communication Disorders, 18*(3), 139-151.

Oksenberg, L., Coleman, L., & Cannell, C. (1986). Interviewers' voices and refusal rates in telephone surveys. *Public Opinion Quarterly*, *50*(1), 97-111.

Owren, M., Berkowitz, M., & Bachorowski, J. (2007). Listeners judge talker sex more efficiently from male than from female vowels. *Perception & Psychophysics, 69*(6), 930-941.

Parris, E. S., & Carey M. J. (1996). Language independent gender identification. *Proceedings of IEEE ICASSP*, 685-688.

Paxson, M. C., Dillman, D. A., & Tarnai, J., (1995). Improving response to business mail surveys. In Cox, B. G., Binder, D. A., Chinnappa, B. N., Christianson, A., Colledge, M. J., and Kott, P. S., (Eds.), *Business survey methods* (pp. 234-255). New York City: John Wiley and Sons.

Pear, T. H. (1931). *Voice and personality*. London: Chapman and Hall.

Peterson, G., & Barney, H. (1952). Control methods used in a study of the vowels. *Journal Acoustic Society of America*, *24*, 175-184.

Picheny, M., Durlach, N., & Braida, L. (1985). Speaking clearly for the hearing impaired: Intelligibility difference between clear and conversational speech. *Journal of Speech Hearing Research, 28,* 96-103.

Rosen, R., Clayton, R., & Rubino, T. (1991). Controlling nonresponse in an establishment survey. *Proceedings of the Survey Research Methods Section of the American Statistical Association,* 587-592.

Ross J. M., Shaffer, H., Cohen, A., Freudberg, R., & Manley, H. (1974). Average magnitude difference function pitch extractor. *IEEE Transactions on Speech and Audio Processing*, *22*, 353-362.

Selvin, H. (1957). A critique of tests of significance in survey research. *American Sociological Review*, *22*(5), 519-527.

Shimamura, T., & Kobayashi, H. (2001). Weighted autocorrelation for pitch extraction of noisy speech. *IEEE Transactions on Speech and Audio Processing*, *9*(7), 727-730.

Singer, E., Frankel, M., & Glassman, M. (1983). The effect of interviewer characteristics and expectations on response. *Public Opinion Quarterly*, *47*(1), 68-83.

Sinibaldi, J. (2010, May 14). *Measurement error in objective and subjective interviewer observations*. Paper presented at the 2010 American Association for Public Opinion Research, Chicago, IL.

Smith, T. (1978). In search of house effects: A comparison of responses to various questions by different survey organizations. *Public Opinion Quarterly, 42*(4), *443-463*.

Smith, T. (1997). Measuring race by observation and self-identification. *GSS Methodological Report No. 89*.

Smith, P. (1979). Sex markers in speech. In K.R. Scherer & H. Giles (Eds.), *Social markers in speech*, (pp. 109-146). Cambridge: Cambridge University Press.

Strand, E. (2000). *Gender stereotype effects in speech processing*. (Doctoral dissertation). Ohio State University.

Strand, E., & Johnson, K. (1996). Gradient and visual speaker normalization in the perception of fricatives. In D. Gibbon (Ed.), *Natural language processing and speech technology: Results of the 3rd KONVENS conference, Bielefeld, October 1996* (pp. 14-26). Berlin: Mouton de Gruter.

Traunmüller, H. (1997). Perception of speaker sex, age, and vocal effort. In Bannert, R., Heldner, M., Sullivan, K., & Wretling, P., (Eds.), *Phonum 4*, (pp. 183-196). Department of Phonetics, Umeå.

Tuomi, S. K. & Fisher, J.E. (1979). Characteristics of simulated sexy voice. *Folia Phoniatrica, 31*(4), 242-249.

U.S. Bureau of the Census. (1972). *Evaluation and Research Program of the U.S. Census of Population and Housing, 1960: Effects of Coders* (Series ER 60, no. 9). Washington, D.C.: U.S. Government Printing Office.

Walker, P., & Cook, D. (1998). Brief Communication: Gender and Sex: Vive la Difference. *American Journal of Physical Anthropology, 106*, 255–259.

Welham, N., & Maclagan, M. (2003). Vocal fatigue: Current knowledge and future directions. *Journal of Voice, 17*(1), 21-30.

West, B. (2010a, May 14). *An examination of the quality and utility of interviewer estimates of household characteristics in the national survey of family growth*. Paper presented at the 2010 American Association for Public Opinion Research, Chicago, IL.

West, B. (2010b). The measurement error properties of interviewer observations and their implications for repairing nonresponse errors. (Doctoral dissertation prospectus). University of Michigan.

Winerman, L. (2005). 'Thin slices' of life. *Monitor on Psychology* , 36, 54. Retrieved March 1, 2011, from http://www.apa.org/monitor/mar05/slices.html.

Yarmey, D. (1995). Earwitness speaker identification. *Psychology, Public Policy, Law. 1*(4), 792-816.