VERBAL PARADATA AND SURVEY ERROR:
RESPONDENT SPEECH, VOICE, AND QUESTION-ANSWERING BEHAVIOR
CAN PREDICT INCOME ITEM NONRESPONSE

by

Matthew E. Jans

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Survey Methodology)
in The University of Michigan
2010

Doctoral Committee:

Professor Frederick G. Conrad, Co-Chair
Professor James M. Lepkowski, Co-Chair
Professor Norbert Schwarz
Assistant Professor Jose R. Benki, Michigan State University
Assistant Professor Frauke Kreuter, University of Maryland
Senior Research Fellow Floyd Jackson Fowler, Jr., University of Massachusetts Boston

Acknowledgements

Many more people and institutions deserve recognition than I have space to provide here, but let this serve as a testament to their support of this research. First, I thank the US Census Bureau (Contract # YA1323-08-SE-0329) and the Charles F. Cannell Fund in Survey Methodology, both of which funded this research and, quite literally made it possible. Second, I thank my dissertation committee members, each of whom offered advice and guidance at key points throughout the process. My co-chairs Fred Conrad and Jim Lepkowski deserve special thanks for hours of weekly consultations; for providing helpful ideas, but also acting as sounding board for my own; for challenging me to design the "perfect" dissertation, but also understanding when the data available would suffice. Third, this dissertation represents a culmination of training and education made possible by the structure and flexibility of the Michigan Program in Survey Methodology (MPSM). Each of the program's faculty contributed to this project in their own "latent" way.

Recordings of telephone interviews were provided by Richard Curtin and Rebecca McBee-Donello at the Survey of Consumers. A dissertation isn't a dissertation without data. Thanks. A fine team of undergraduate coders produced the data that are analyzed in this dissertation. They are Maya Burns, Alex Dopp, Zach Hartley, Clare Levioki, Lian Liu, Miyuki Nishimura, Melinda Mosher, Travis Pashak, Katie Singer, and Jenna Stein. I only hope the experience was as valuable to them as their commitment and work was to me.

MPSM graduate students rated the survey for sensitivity and complexity, thus established a key part of the study's design. Thanks guys!

Several people outside of MPSM also have my sincere gratitude, specifically two consultants at the Center for Statistical Consultation and Research (CSCAR) at the University of Michigan. Laura Klem and Joe Kazemi each provided several hours of consultation on the analysis of these data. They provided a range of advice from sage wisdom to specific SPSS code, each priceless in its own way.

I want to also thank my wife Laura Clary (of about one year as this goes to print). I relied on her emotional and intellectual support regularly throughout the project, and plan to do the same for her soon. Reflecting on the second-hand dissertation experience, she recently said "You weren't all as crabby as you said you'd be!"

Many other students, friends and colleagues have also contributed in more or less tangible ways, including discussions of the ideas and data presented here, and one-off consultations. Although there is not enough space to list each of these people individually, I think you know who you are. Thanks!

Table of Contents

List of Figures

List of Tables

List of Appendices

Abstract

VERBAL PARADATA AND SURVEY ERROR:
RESPONDENT SPEECH, VOICE, AND QUESTION-ANSWERING BEHAVIOR
CAN PREDICT INCOME ITEM NONRESPONSE

by

Matthew E. Jans

Co-Chairs: Frederick G. Conrad and James M. Lepkowski

Income nonresponse is a significant problem in survey data, with rates as high as 50%, yet we know little about why it occurs. It is plausible that the *way* respondents answer survey questions (e.g., their voice speech, and question-answering behavior) can predict whether they will provide income data, and reflect the psychological states that produce this decision. Five questions each from 185 recorded interviews conducted by the Surveys of Consumers were selected. One was the annual household income question. Exchanges between interviewers and respondents were transcribed and coded for respondent speech and question-answering behavior. Voice pitch was extracted mechanically using the Praat software. Speech, voice, and question-answering behaviors are used as verbal paradata; characteristics of the survey process that are not captured by default. Verbal paradata are hypothesized to reflect respondents' affective and cognitive states, which then predict income nonresponse. It was hypothesized that indicators of respondent affect (e.g., pitch) and cognitive difficulty (e.g., disfluency) would be affected by sensitive and complex questions differently, and would predict whether respondents provide income in a dollar amount, a bracketed range of values, or not at all. Results show that verbal paradata can distinguish between income nonrespondents and respondents, even when only using verbal paradata that occur before the income question. Income nonrespondents have lower affective involvement and express more negativity before the income question. Bracketed respondents express more signs of cognitive difficulty. Income nonresponse is predicted by behavior before the income question, while bracketed response is predicted by indicators on the income question itself. Further, question characteristics affect respondent paradata, but largely in unpredicted ways. There is evidence for psychological resource and conversationality mechanisms through which respondents *reduce* verbal paradata when questions are demanding, rather than increasing it as signs of trouble. The results have implications for theory of income nonresponse, specifically the role of question characteristics and respondent paradata in understanding what subjective psychological states respondents are experiencing when they answer survey questions, and how those states predict whether income is reported. There are also potential extensions to interviewer training and design of interventions that could produce more complete income data.

Chapter 1

Income Nonresponse, Question Characteristics and Respondents' Verbal Paradata

1.1 Dissertation Overview, Goals, and Conceptual Framework

Income data quality (item nonresponse and measurement error) is an ever-present risk to overall survey data quality. When survey items ask about income (or other sensitive or cognitively complex topics), the risk to item-level data quality is particularly high. Additional cognitive and social processes are likely at work when answering sensitive and complex questions, like income, beyond those involved in answering nonsensitive and noncomplex questions. Respondents' psychological states cannot be observed directly, so observable features of the survey response that represent those psychological states must be relied upon. Observable features of survey responding include question-answering behavior (e.g., a request for clarification), speech characteristics (e.g., pauses, fillers, and other disfluency), and acoustic qualities of voice (e.g., pitch).

Income questions pose specific data quality concerns. They can be both sensitive and complex, and often have high levels of nonresponse (Moore, Stinson, & Welniak, 2000; Moore & Loomis, 2001; Moore, 2006; Yan, Jans, & Curtin, 2006). This dissertation evaluates observable indicators of respondents' psychological states, specifically *affect* and *cognitive difficulty* that can be measured in their question-answering behavior, speech, and voice. These indicators of psychological states are measured as respondents answer five questions, including one about annual household income, that vary in sensitivity and cognitive complexity. The goals are to understand 1) how survey questions with varying levels of sensitivity and cognitive difficulty influence how respondents answer questions, and 2) how these verbal paradata relate to income nonresponse.

1.2 Income Data Quality, Respondent Psychology, and Paradata

1.2.1 The Problem of Income Nonresponse

Questions about personal and household income have had the attention of survey methodologists for decades due to their recognized potential for poor data quality (Shih 1983; Bell, 1984). Poor-quality data from self-reported income can be found in multiple components of survey error, such as measurement error, coverage error, and so forth (Groves, Fowler, Couper, Lepkowski, Singer, & Tourangeau, 2004; Groves, 1989). Specific causes of error can be found in the form of the question (Schuman & Presser, 1977), the specificity of response required (Moore, & Loomis, 2001), the mode of administration (Tourangeau & Smith 1996), and characteristics of the respondent (Bell, 1984; Nicoletti & Peracchi, 2001; Riphahn & Serfling, 2005). The error component explored in this dissertation is item nonresponse (i.e., item missing data), and causes of interest center around respondents and how they answer or don’t answer an income question. Item nonresponse occurs when respondents agree to participate in a survey, answer some of the survey questions posed to them, but do not answer one or more questions within the survey. In interviewer-administered surveys, a respondent generally must make a verbal indication that they do not want to answer a question before the interviewer will move on to the next question. This verbal behavior may contain useful information for predicting future income nonresponse.

The presence of nonresponse is a serious problem for statistical estimation from survey data, regardless of its causes. Like unit nonresponse error, item nonresponse error is a function both of the item nonresponse rate, and the difference between respondents and nonrespondents (Groves 1989; Groves et al., 2004; Groves, 2006). A sufficiently high item nonresponse rate, combined with a sufficiently high difference between respondents and nonrespondents will produce nonresponse error in a statistic (e.g., a mean, proportion, or regression coefficient) based on that item. Nonresponse error can also be conceived of as a stochastic process involving the correlation between a respondent's true value on a survey question and their propensity to report their value on a survey questionnaire.[1] If the difference (or correlation) is systematic (e.g., the same direction and degree) across all conceptual replications of the survey under identical essential survey conditions, nonresponse bias results. If the difference (or correlation) varies across conceptual replications, the error is considered a variable error, or nonresponse variance. In either case, the item nonresponse rate is one important factor in understanding item nonresponse bias, even though rates alone do not determine nonresponse (Groves, 2006). Even in the absence of item nonresponse bias (e.g., item missing data that are completely at random; Little & Rubin, 2002), higher item nonresponse rates pose problems for the analysis of survey data, reducing the complete-

[1] $Bias\ \overline{y}_r = \frac{nr}{n}(\overline{y}_r - \overline{y}_{nr})$ and $Bias\ \overline{y}_r = \rho(y,p)/\overline{p}$ are the deterministic and stochastic formulae for nonresponse bias respectively, where y is income, $\overline{y}_r$ is the mean income of sample respondents, $\overline{y}_{nr}$ is the mean income of nonrespondents, $\frac{nr}{n}$ is the nonresponse rate, and p is the response propensity.

case sample size for any analysis that uses the item with missing data. Smaller sample sizes lead to increased standard errors of estimates, even if the estimates are unbiased. Thus, reductions in nonresponse rates are an important goal, even if bias is not affected. In an effort to understand what causes income nonresponse rates, this dissertation deals only with nonresponse and its causes, recognizing that it is only a piece of the income data quality problem.

1.2.2 How Often Does Income Nonresponse Occur?

Income nonresponse is a major problem in survey research and quantitative social research more broadly. Income nonresponse rates can be as high as 50% for interest and investment income (Juster, Smith, & Stafford, 1999; Moore et al., 2000). While reports of annual salary and wages may have low rates of missing data as low as 3-7% for some months of the Surveys of Consumers (Yan et al., 2006), other research finds item nonresponse rates for annual income (overall, and wages and salaries independently) that are as high as 25-30% (Atrostic & Kalenkoski, 2002; Yan et al., 2006). Income nonresponse rates often fall within the range of 10-25% (Atrostic and Kalenkoski, 2002; Battaglia, Hoaglin, Izrael, Khare, & Mokdad, 2002; Dixon, 2005; Juster and Smith 1997; McGrath, 2005; Moore et al., 2000; Olson et al., 1999).

A unique facet of income nonresponse is highlighted by comparing income item nonresponse rates with item nonresponse rates on other survey questions. Survey items typically have nonresponse rates in the range of 1-4% (de Leeuw, 1992; de Leeuw, Hox, & Huisman, 2003; Bradburn, Sudman, & Associates, 1979; Tourangeau, Rips, & Rasinksi, 2000). Few items produce higher nonresponse rates than income items, and those that do contain uniquely sensitive subject matter, such as sexual behavior and drug use (Tourangeau et al., 2000). The fact that income items rank with other sensitive questions on item nonresponse rates suggests that sensitivity is likely a factor in producing income nonresponse. Many surveys, however, ask very few sensitive questions, and thus the largest missing data problem for those surveys will likely come from questions about income. Income items are clearly problem items for many researchers.

This dissertation deals only with data from a telephone survey, but mode effects on income nonresponse help frame the current study in a broader context of income nonresponse rates and error. It is clear that income nonresponse rates can differ across modes of data collection, but findings are mixed. Of those studies that explicitly compare rates across modes, some find no differences on financial and income questions (DeLeeuw, 1999), while others find differences by mode, with more nonresponse in telephone surveys (Kormendi, 1988; Körmendi & Noordhoek, 1989; Schraepler, Schupp, & Wagner, 2006). Studies that explore income nonresponse rates through nonexperimental comparisons find the same mixed pattern. For example, the Current Population Survey, which is a phone and in-person survey, has income nonresponse rates between 14% and 27% (Atrostic & Kalenkoski, 2002; Moore et al., 1999; Dixon, 2005). Similarly, the National Immunization Survey (a phone survey) has documented income nonresponse rates from 17% to 32% (Olson et al., 1999) and the Consumer Expenditure Quarterly Survey (an in-person survey) finds income nonresponse rates between 19.9% and 35%.

Income nonresponse rates also differ across groups of respondents and across values of income. Increased income nonresponse is found in older respondents, and

White respondents compared to younger and non-White respondents (Bell, 1984; Riphahn & Serfling, 2005). Further, adding an open-ended income question to a previously bracketed income series obtains additional responses from lower-income respondents, suggesting that income estimates based on brackets alone may be positively biased. The finding that lower income respondents respond at higher rates when an open-ended income question is added to a bracketed one suggest that question format and true income value can interact to produce income nonresponse (Bell, 1984). From this finding, however, it is not clear whether it was the open-ended response, the second chance to provide income, or the order of the two response formats that brought in additional responses. Self-employed respondents produce more income nonresponse, though this does not seem to bias income estimates (Nicoletti & Peracchi, 2001). Similarly, there is a tendency for respondents not employed full-time to provide less complete income data than those employed full-time (Riphahn & Serfling, 2005).

1.2.3 What Causes Income Item Nonresponse?

Income item nonresponse is a special type of item nonresponse that involves content that can be both sensitive and cognitively challenging. The causes of item nonresponse, generally speaking, are not completely understood but a three-cause model has been proposed to explain why respondents choose not to respond to individual survey items (Beatty & Herrmann, 2002). According to this model, the causes of item nonresponse for any given respondent could be one or more of the factors of cognitive state, adequacy judgment, and communicative intent. Cognitive state and adequacy judgments (i.e., the level of detail required to answer and the level of detail available to the respondent) are subsumed under cognitive reasons for item nonresponse in this model. Communicative intent in this model is synonymous with motivation, under which Beatty and Herrmann include sensitivity, cognitive burden, and conflict of interest. It is particularly strange, conceptually speaking, that “cognitive burden” is classified as “communicative intent” and not “cognitive state”. While Beatty and Herrmann’s taxonomy of item nonresponse causes may not be entirely inclusive or fully specified[2], findings derived under this model are helpful to understanding item nonresponse. Testing their model in a paper-and-pencil self-administered survey of undergraduate students, Beatty and Herrmann (2002) found that cognitive reasons, rather than motivational ones were cited for item nonresponse. This is not surprising, because the survey was short, the respondents were “captive” (i.e., students in a classroom), and items tested were primarily non-sensitive, asking about memory for events and dates. Reasons for nonresponse (e.g., retrieval difficulty, adequacy judgment, etc) were all measured through self-report as well, and respondents’ abilities to report accurately about these cognitive processes are questionable (Ericsson & Simon, 1980; Ericsson & Simon, 1993). Further, the self-administered mode may not generalize to social and psychological nonresponse causes that are activated by an interviewer’s presence. It is likely that the psychological causes of item nonresponse differ when an interviewer is present (de Leeuw, 1992; de Leeuw, Hox, & Huisman, 2003).

Many cognitive processes are involved in answering survey questions, and the specific stages and processes depend on the requirements of the question. With each

---

[2] For example, the model includes cognitive factors under both “cognitive state” and “motivation” (in the form of cognitive burden of retrieving the required information). Moreover, the “motivation” factor also seems somewhat conceptually overly-inclusive, including cognitive, social, and affective reasons for nonresponse. No category is specified for “motivation to conduct a memory search” specifically, rather their term “motivation” seems to be a place-holder for all social and affective (e.g., non-cognitive) reasons for nonresponse.

process, there is a possibility for nonresponse error. Irrespective of how a respondent feels about a survey question or topic, a respondent must comprehend the question (e.g., words, meaning, goal), conduct a memory search (if the question asks for facts or events), create a response, and decide how to report the retrieved or calculated value within the requirements of the survey question (e.g., response options). Most models of survey responding propose some version of this basic psychological response model (see Tourangeau et al., 2000 for a review). Affective components of response (i.e., how the respondent feels while answering, and the effect of feeling on response quality) are generally excluded or only loosely specified. This dissertation is motivated by the perspective that affect and cognition both operate simultaneously to produce survey data.

1.2.4 Affect and Cognition as Forces in Survey Error

The primary motivation for the study of affect and cognition in survey responding is that psychological states are made up of two general classes of states (affect and cognition) that are involved in survey response. Affect[3] and cognition are two wide-reaching psychological dimensions that influence survey response. Indeed voluminous work has been done on affect and cognition independent of survey methodology (Christianson, 1992; Forgas, 2001; Lewis, Haviland-Jones, & Barrett, 2000; Schwarz, 2000; Schwarz & Clore, 2007), while the cognitive aspects of survey methodology

---

[3] Affect and emotion, both considered to be “feeling states” (Schwarz & Clore, 2007), can be distinguished from each other on the basis of several factors. Emotions tend to be best described as discrete feeling states, with clear referents, acute onset and rise, and shorter duration. For example, feelings of frustration by heavy traffic on one’s commute may be intense in the moment, but subside once at home. Affect is better described as a regular component of feeling; while emotions can come and go, affect is ever-present. It also often does not have a clear referent or cause, has gradual onset, and has longer duration. It is better described in terms of valence (positive or negative) and intensity, rather than discrete states, as is the case for emotions. For the purposes of the dissertation, the term affect will be used generally to refer to all feeling states, though affect also is the primary feeling state of interest.

(CASM) movement has brought these psychological factors into the field of survey methodology (Sirken, Herrmann, Schecter, Schwarz, Tanur & Tourangeau, 1999). To the degree that psychological states consist of affective components and cognitive components, we should seek to understand if one or the other is more influential on respondents' propensity to provide good data. Knowing what components of a respondent's psychological state are responsible for (or at least predict) error can be a first step toward developing methods to reduce it. These psychological states also map closely to characteristics of survey questions, with affect mapping most directly to a question's sensitivity and cognition (cognitive difficulty) to a question's cognitive complexity. Error produced by the sensitivity and complexity of survey items will thus likely arise in the affective and cognitive responses to the questions.

Affect, Cognition and Their Interaction

The independence or correlation of affect and cognition, and their causal order, have been debated in psychology since the founding of the field (see Ellsworth, 1994 for a discussion of the debate from William James' original propositions; also Diener & Emmons, 1984; Russell, 1980; Schachter & Singer, 1964; N. Schwarz & Clore, 2003; Schwarz & Clore, 2007; Zajonc, 1980). Underlying this discussion is an explicit understanding that affect and cognition are shaped by a person's physiological arousal, which can be influenced by stimuli in the environment, as well as internal psychological states. Exploring the relationship between affect and cognition helps motivate the relationship of voice and speech with psychological states, psychological states with each other, and those states with income nonresponse. Without this broader context, it is tempting to think of affect and cognitive difficulty as completely independent. A brief review of theory and research from social and cognitive psychology help describe the nature of these phenomena.

The order of affect, cognition, and physiological response originally proposed by William James led from an environmental stimulus, to an interpretation of that stimulus, to a physiological response, to an emotion. This was a revision of the contemporarily held belief that after interpretation of a stimulus, an emotion was experienced that leads to a physiological response (Ellsworth, 1994). Schachter and Singer (1962) provide evidence that physiological arousal interacts with the social context to influence feelings about stimuli. Individuals will interpret physiological states differently (i.e., apply affect labels to them) depending on the attributions available for the psychological arousal. For example, when physiological arousal is increased by a drug, and a potential attribution is present in the social context (e.g., an irritating partner in a group interaction) people are likely to attribute their arousal to anger at the partner (i.e., the social context) when they are unaware that their arousal had been heightened by the drug (Schachter & Singer, 1962). Physiological arousal influences how people interpret social experiences, and respondents may do the same with surveys, placing feelings caused by something else (e.g., the weather) onto the survey or the interviewer (Schwarz & Clore, 1983).

The series of subjective responses to environmental stimuli, which includes a physiological reaction, an affective reaction, and a cognitive reaction has changed over time as theory of emotion has developed. In contemporary theory, affect is placed as the first response to a stimulus, followed by an interpretation of that affect and then a physiological response. Under this model, people experience the world (including survey questions) affectively first, then cognitively. This has been called a “feelings as information” perspective on affect and cognition (Schwarz & Clore, 2007). From this perspective, respondents may use how they feel about the survey experience or about providing income as relevant to the decision whether to report income.

Theory and research on the relationship of affect and cognition is still active within psychology. One model of affect essentially suggests that affect states are more connected than different, and they can be placed on a circular continuum (Russell, 1980). This circumplex model of affect has been critiqued by those who hold that affect and emotion are independent (Diener & Emmons, 1985). In this dissertation, connections between (and independence of) affective states will not be explicitly tested, and a general affect arousal perspective will be taken, in which respondent verbal paradata are expected to signal heightened affective arousal in respondents. Although *specific* affective states (e.g., anxiety, nervousness, joy) may be operating as respondents answer survey questions, it is not clear that these can be easily measured from audio recordings (Bacharowski, 1999), or whether they are helpful to understanding the role that “everyday affect” plays in making decisions about survey responding (Schwarz, 2000). A more explicit theoretical exploration in this dissertation will focus on the relative role of affective and cognitive processes (as measured by verbal paradata) in predicting income nonresponse, and the degree to which these processes are affected by characteristics of survey questions.

It is clear that both affect and cognition operate together to lead to decisions (Schwarz, 2000; Schwarz & Clore, 2003, Schwarz & Clore, 2007), such as the decision to provide income data in response to an interviewer’s request. In some circumstances, affect may precede other psychological actions, and in others, cognition (or cognitive difficulty) will come first. Affect that influences decision processes can come from different sources. Affect related to an object about which a decision is to be made has been referred to as integral affect, because it is integral to the decision object (Schwarz & Clore, 2007). Incidental affect on the other hand, is not directly related to the decision object (i.e., the mood that the respondent happens to be in when the decision has to be made).

Affect seems to have a larger role in decision processes than originally thought (Cacioppo, Petty, Feinstein, & Jarvis, 1996; Fabrigar & Petty, 1999; Hox, de Leeuw, & Vorst, 1996; Schwarz & Clore, 2003; Schwarz & Clore, 2007), and thus likely has a large role in the psychology of survey response, including income item nonresponse. Respondents' affective reactions to specific questions or other essential survey conditions (e.g., mode, order of questions, interviewer characteristics and behavior) could be called integral affect (Schwarz & Clore, 2007), because the survey itself is the object requiring decision and is producing the affective response. Sensitivity of a topic or an interviewer characteristic (e.g., things that are in the control of the researcher) as well as the respondent's true value (i.e., a facet not under the control of the researcher) can be thought of as integral affect. This dissertation will evaluate integral affect that predicts income item nonresponse, specifically respondents' affective reactions to survey questions. Incidental affect, the affect that is experienced by the respondent but not directly related to any facet of the survey (i.e., the mood that the respondent happens to be in when they take part in the interview), can also lead to item nonresponse. Incidental affect will not be studied in this dissertation. It is worth acknowledging these different types of affect if only to recognize that affect states and processes are due to more than the “stimulus” that the survey provides to the respondent.

In addition to affective responses to survey questions that might cause item nonresponse, respondents may not answer income questions because they are too difficult (Juster & Smith, 1997). Evidence of cognitive difficulties in survey responding has been documented on questions that pose challenges for respondents (Schober and Conrad, 1997; Conrad and Schober, 2000; Conrad et al., 2008; Draisma & Dijkstra, 2004; Ehlen, Schober, & Conrad, 2007; Schober & Bloom, 2004). Contemporary models of survey responding (e.g., Tourangeau et al, 2000) and research sparked by them seem to have a stronger focus on cognition than affect. It is clear that how respondents cognitively process survey questions (independent of their feelings about them) affects response. The CASM movement has produced much research on the psychological and social dimensions of data quality (Sirken et al., 1999). Yet these studies tend to be more about “thinking” and less about “feeling” when answering survey questions. Research and theory from social and cognitive psychology have explained the relationship between cognition and emotion (Zajonc, 1980; Schwarz & Clore, 2007), and emphasize that thoughts often flow from affective judgments or feelings about stimuli in the environment (e.g., survey questions or interviewer behavior). Applying this research to survey methodology shifts focus from what respondents might be thinking to what they might be feeling when deciding to respond to a survey request, answer a question, or while reporting an answer. This dissertation explores the dual roles of affect and cognition in respondents’ verbal reports to survey questions.

1.2.5 How to Study Income Nonresponse and Its Causes

Many variables can describe the data collection process, but only a few of these are actively measured in common practice. Collectively, these variables have been labeled *paradata*, and have been applied to the measurement and prediction of survey error (Couper, 1998). It is helpful to distinguish between *potential* paradata and *actual* paradata. Any variable that describes the data collection process can fall under the rubric of actual paradata, if it is recorded. Yet, many potential paradata go unrecorded in standard survey data collection. Even those that are collected as administrative data during the data collection process are rarely analyzed. Examples of potential paradata include date and time of the interview, number of contacts required to complete the interview, or interviewer or respondent characteristics. At a finer level of detail, potential paradata can include data entry key strokes made by interviewers working in a computerized instrument, moment-by-moment exchanges between interviewers and respondents, or the millisecond-by-millisecond order and quality of respondent speech and voice within individual spoken utterances[4]. These micro-level paradata have potential to provide measures of psychological processes involved in income responding and survey error more generally (Draisma & Dijkstra, 2004; Maynard et al, 2004; Schober & Bloom, 2004; Conrad, Schober, & Dijkstra, 2008). Response latencies have been used heavily in psychological research as measures of cognitive processes (Draisma & Dijkstra, 2004; Bargh, Chaiken, Govender, & Pratto, 1992). Analyses of voice and speech also have a long history of fruitful research in cognitive and social psychology,

[4] Utterances are generally defined as the smallest unit of speech that holds semantic comment. Thus multiple utterances can be contained within one conversational turn. The distinction between utterances and turns and their application to the dissertation will be addressed more thoroughly in Chapter 2.

and the study of affect (Bachorowski, 1999). These uses will be explored further in this dissertation.

Only if potential paradata are recorded in some way (e.g., coding of audio recordings as in this dissertation) do they become actual paradata, comprising a data set that can be linked to survey outcomes and analyzed.[5] The ultimate utility of paradata depends on the degree to which they can predict and explain survey error (e.g., variances and biases stemming from different parts of the measurement and estimation process). To the degree that these paradata also represent psychological states of respondents, we can link psychological states to survey error as well. This dissertation is specifically focused on paradata in the form of respondent voice, speech, and question-answering behavior, their production as a reaction to item sensitivity and complexity, and their ability to predict item nonresponse on an income question. Paradata consisting of respondent speech, voice, and question-answering behavior will be collectively referred to in the dissertation as "verbal paradata".

For interview modes, in which respondents answer questions verbally, sensitive and complex survey questions may elicit evidence of the cognitive and affective response process in the spoken words of respondents. The psychological state experienced by a respondent comprehending, retrieving, and editing a response to a sensitive question may be different from a respondent taking the same cognitive steps to answer a cognitively complex one. Cues to the nature of these psychological states may be evident in respondents' voice and speech (e.g., respondent verbal paradata).

[5] Throughout the rest of the dissertation the term "paradata" will be used to refer to potential and actual paradata. The distiniction, if one needs to be made, should be clear from the context.

One way to measure and study income nonresponse is to use facets of the answer itself. A distinction is sometimes made between respondents who refuse to answer the question (e.g., “I don’t want to answer that” or “I won’t give that out”) and respondents who say that they do not know or do not have the required information (e.g., “I don’t know” or “I couldn’t tell you that without looking it up”). Distinguishing between refusal and don’t know responses is helpful in that it provides an additional characteristic of the income nonresponse. This characteristic may or may not reflect the of the income nonresponse. The utility of this distinction for understanding ultimate reasons for nonresponse depends on the degree to which the labels “don’t know” and “refusal” accurately capture respondents reasons for not responding. Some authors suggest that “don’t know” responses can be refusals worded in a more polite manner (Moore & Loomis, 2001). The respondent has the information required to answer the question, but is unwilling to report it, and answers “I don’t know”. Don’t knows and refusals alone do not capture verbal facets of the response, and thus are likely to be weak proxy measures for psychological states of respondents. This study does not predict differences between don’t knows and refusals on the income question. Rather, these labels are considered question-answering behavior, and are used as predictors of income nonresponse.

Current research and theory on income nonresponse motivate more in-depth and extensive studies of exactly how respondents answer income questions (de Leeuw et al., 2003) and the role of cognition and affect more generally (Beatty & Herrmann, 2002; Moore, Stinson, & Welniak, 2000).

Affect and Acoustic Properties of Voice

A body of research with potential application to survey methodology shows that affective states relate to differences in voice production (Bachorowski & Owren, 1999; Leinonen, Hiltunen, Linnankoski, & Laakso, 1997). While it is difficult to link specific discrete emotions to voice (e.g. anxiety, fear, depression, sadness), the link between voice and affect intensity (high or low) and valence (positive or negative), is much easier to show. Increased nonspecific emotional arousal is associated with increases in voice pitch. For example, increased pitch is associated with emotional states of joy and anxiety, both of which are also characterized by physiological arousal. It follows from this finding that pitch is not a good indicator of discrete affective states, and perhaps even affect valence. However, it is at least a reasonable reflection of a respondent's subjective affective and physiological arousal (Bachorowski, 1999).

Given that acoustic properties of voice have been linked to affective states in speakers, they can be used as indicators of unmeasured (latent) affective states of respondents as they answer survey questions. If findings from psycholinguistics are applicable to the survey interview context, then acoustic qualities of respondents' voices should be indicative of their affect, where increased affect intensity (specifically negative intensity related to anxiety or frustration) would be expected on questions that are sensitive. Fundamental frequency (pitch), intensity (volume), jitter (vocal fold vibration frequency variability) and shimmer (vocal fold vibration amplitude variability) may all be related to affect (Bachorowski, 1999). Measurement of voice via telephone is limited to pitch as it is the most robust measure of this group. The use of pitch as an acoustic quality has precedent in survey methodology, but it has been used primarily to study unit nonresponse, where the role of the interviewer's pitch is under question (Groves, R. M., O'Hare, Barbara, C., Gould-Smith, D., Benki, J., & Maher, P., 2008; Oksenberg & Cannel, 1988; Van der Vaart, Ongena, Hoogendoorn, & Dijkstra, 2006). No research was found using respondent voice as a predictor of survey error, or that looks at income nonresponse.

The literature on voice and affect supports predictions that variability in pitch and increases in pitch should indicate increases in affect (i.e., heighted affect arousal). Such changes in affect should be driven by the sensitivity of survey questions, and so it is expected that sensitive questions will lead to increased pitch and pitch variability.

Cognitive Difficulty and Speech

Apart from acoustic properties of voice, spoken words have cadence and pacing (e.g., pauses, rate of speech, fluency of speech), which collectively make up a body of verbal behaviors that will be referred to simply as speech. Speech is conceptually distinct from, though sometimes behaviorally related to acoustic properties of voice. One can imagine a monotone, mono-volume voice (little variation in acoustics) that is highly disfluent (many pauses, stammers, fillers like *um* and *uh*, and varying rates of speech). For this speaker the speech production facility is distinct from the acoustic facility. One can also imagine the opposite, a speaker who has much acoustic variation and is completely fluent. Voice and speech characterize conceptually independent characteristics of spoken word. This distinction is important, not only because acoustic properties and speech production define different facets of spoken word (Kent, 1997), but also because those facets have been linked to different psychological states (Bachorowski, 1999; Bargh et al., 1992).

While affect has been tied primarily to acoustic characteristics of voice (e.g., pitch), cognitive difficulty is linked more clearly to speech, or the way in which words are produced by the speaker. Some of that evidence comes from research in survey methodology (Conrad et al., 2008; Draisma & Dijkstra, 2004; Ehlen, Schober, & Conrad, 2007; Schober & Bloom, 2004). Higher rates of pauses, response latencies, fillers (e.g., *um*'s and *uh*'s), and repairs have all been linked to inaccuracies in survey responses.

In addition to speech, which might be considered a "personal style" of verbal responding, other more discrete types of response behavior and speech acts have been linked to difficulty with survey questions. For example, reports, in which respondents state facts about their personal situation relevant to the survey question without directly answering the question, are found more frequently in complicated questions (Schober & Bloom, 2004). Verbal paradata like these will be referred to as question-answering behavior.

Conversation and the Survey Interview

The survey interview is often viewed as a technical, official interaction that happens to have cognitive and social components. It can also be viewed as a social interaction that happens to be official (Maynard et al., 2002; Suchman & Jordan, 1990; Suchman & Jordan, 1991; Cassell & Miller, 2008; Cassell, Gill, & Tepper, 2007). The affective component of interviewer-respondent interactions and psychological response processes has received less attention than the cognitive aspects of response. A few conceptual and empirical investigations have explored "affect-like" dimensions of data quality (e.g., likeability, etc, see Oksenberg, Coleman, & Cannell, 1986), but these have been limited to hypothetical situations in which respondents' feelings about personal qualities of interviewers are correlated with unit nonresponse. No research has been done that links self-report data quality to *respondent* affect.

1.3 Sensitive questions, Complex Questions, Verbal Paradata and Income Reporting

Because income questions can be both sensitive and cognitively complex, using those question characteristics to help understand income nonresponse may be advantageous. Sensitive questions are hypothesized to elicit affective states such as embarrassment, anxiety, fear, and discomfort, all of which are feeling states associated with the sensitive or threatening nature of the question. Sensitive questions are those that ask about topics that are personal or private, or put the respondent at a risk if true answers are disclosed. They may be factual or attitudinal. Cognitively complex survey items are those that are difficult for respondents to answer due to the number and complexity of mental operations required, such as mathematical calculations, memory searches, or judgments. Complex items often ask for factual information (or information that potentially has an exact, factual answer, even if factual answers aren't required for the response), which is part of what makes them challenging for respondents. The specific qualities of sensitive and complex items will be reviewed in more depth.

1.3.1 Sensitive Questions

There is no unanimously agreed-upon definition of what constitutes a sensitive question. Current theory on sensitive questions proposes three types of sensitivity: social desirability, invasion of privacy (intrusiveness), and risk of disclosure (Tourangeau, Rips, & Rasinksi, 2000; Tourangeau & Yan, 2007). Social desirability deals specifically with the social interaction between the interviewer and respondent. Social desirability explanations argue that respondents have a need to present themselves in a positive way to interviewers. They accomplish that through their answers to survey questions for which one response could be seen as a more positive behavior or characteristic than another. In psychological literature, socially-desirable responding has been tied to personality dispositions (Crowne & Marlowe, 1960).

A social desirability explanation of income data quality would argue that respondents will consciously modify their income reports from their true value in order to bring them toward their perception of the average income, or in line with what they might assume the interviewer earns. This would result in measurement error. While there is strong evidence that social desirability is the cause of under-reporting of some behavior or characteristics like library card ownership, voting, abortions, sexual behavior, and drug use (Tourangeau & Smith, 1996; Tourangeau, Smith, & Rasinski, 1997; Traugott, 2008), it is not clear whether the same mechanism is responsible for item nonresponse on sensitive questions. If a respondent feels social pressure to not report their income accurately (to the best of their ability), they either have the option to misreport (e.g., reporting a lower or higher income to bring their report closer to that of the interviewer's assumed income), or to not report at all. From the literature on sensitivity and item nonresponse, it is not quite clear what would motivate respondents to misreport income rather than refusing to report it (Tourangeau & Yan, 2007), but the mechanisms for each may be similar

Another explanation of item sensitivity describes it as a matter of intrusiveness, in which cultural norms define which topics are polite to discuss with strangers. Under this definition of sensitivity, a respondent might refuse to report income because he or she feels that it is none of the interviewer's (or researcher's) business, or that it is rude for the interviewer to have asked about income. It is not clear whether respondents would provide an incorrect response to questions that are perceived to be intrusive (measurement error) or simply refuse to respond (nonresponse error potential).

A final definition of sensitivity relates to the perceived risk of disclosure. Under this definition, respondents will withhold or modify information that they worry will be disclosed beyond the research project (e.g., to a third party)[6]. If respondents perceive that their income values will be viewed by other individuals, or somehow shared with another agency (e.g., the IRS), they may be likely to withhold that information or modify that information so that it is not their true value.

It is possible to distinguish between items that are sensitive due to their content, and those that are only sensitive if a respondent has a particular true value. Questions about drug and alcohol use, sexual behavior, income and finance, some health problems

---

[6] The concept of disclosure risk in survey methodology tends to be used to discuss disclosure outside the project, such as a data security breech from outside the institution, or the ability of a third party (i.e., not the interviewer, principal investigator or project staff) to be able to link the respondent's survey responses with their identity (Couper, Singer, Conrad, & Groves, 2008). However, from the perspective of the respondent, the definition of disclosure might be broader than that, and involve a hierarchy of "disclosure distance". If the interviewer is one step removed from the respondent's personal privacy (i.e., answering income is "disclosing" income to the interviewer), then other interviewers on the project staff could be considered two steps, other project staff (project managers and PI's) could be considered three steps, and beyond the project (i.e., a "third party") a fourth (and most severe) step. Assuming that most respondents do not fully understand survey project management systems and data structures, they may assume that ONLY the interviewer is seeing their data, and not realize that other project staff *will be able to* link their responses to their contact information and identity, even if this is not done in practice. Indeed, at least one respondent in the University of Michigan/Reuters Surveys of Consumers believed that the interviewer knew her name and address although she was selected via random digit dial. When dissuaded of that, the insightful respondent pointed out that the interviewer could attach her name and location by using a simple "reverse search" function available on the Web. It becomes clear that risk of disclosure is a complicated subject and includes more levels than just third-part intrusion and security breech.

(e.g., genital, urinary or gastro-intestinal problems) are often considered sensitive due to their content alone. The notion is that these are private topics that are no one's business but the respondent's. Another way to think about sensitive questions is that questions themselves are not sensitive, but responses are sensitive. Thus, sensitivity depends on the respondent's true value with respect to the question. A question on colorectal cancer may not be sensitive to someone who does not have the disease, but sensitive to someone who does. It may also be sensitive if, for example, the respondent's father has colorectal cancer. A further distinction of item sensitivity argues that the specific context of the data collection is what creates the sensitivity (and thus potential for error in reporting). Facets of the data collection context might include the mode of data collection (e.g., whether another person is present or the answer has to be spoken out loud) or the race or sex of the respondent (e.g., a question about racial attitudes may be sensitive if the interviewer is of a different race than the respondent, but not sensitive if they are the same race). This dissertation will work primarily from the assumption that question content (e.g., income, drug use, sexuality) defines sensitivity. Question topics are used as proxies for sensitivity, but respondent true values are rarely if ever available to be taken into account.

Item sensitivity can be evaluated in a number of ways. With some methods, establishing sensitivity is exogenous to the survey response process. Having survey design experts or a group of potential respondents (but not actual respondents) rate the sensitivity of survey items can be a helpful way to classify items on causal dimensions of interest (e.g., sensitivity). These items and their classifications can then be evaluated on facets of the survey response process after the instrument is fielded (e.g., behavior or other paralinguistic codes, or response distributions and item nonresponse rates).

Other more endogenous methods of assessing item sensitivity based on survey data can also be used, but conceptual problems arise with these. For example, nonresponse rates can be used as proxies for sensitivity, assuming that more nonresponse is indicative of sensitivity. One problem with this approach is that qualities of the data are only a proxy for the true concept of interest, which is the respondent's psychological state (i.e., perception of sensitivity) while answering the question. High item nonresponse could be due to other causes (e.g., cognitive difficulty or forgetting) that have nothing to do with sensitivity. Another problem with this approach, particularly when data quality (e.g., item nonresponse) are the outcomes of interest is that it risks setting up a logical loop by which the consequent is defined by the antecedent. [7] If item nonresponse rates are used to define sensitivity, and the effect of sensitivity on item nonresponse is also of interest, it's not quite clear how to conceptually distinguish the causal mechanism from the outcome. Characteristics of responses might be more helpful when they are compared across modes, but it is still important to consider other ways of determining sensitivity, since the respondent's perception of sensitivity is not directly measureable.

Self reports of sensitivity (e.g., "how comfortable or uncomfortable were you answering questions about income") have their own inherent measurement problems if the construct of interest is respondent anxiety. The ability to accurately attend to and report one's own cognitive processes is limited (Erikson & Simon, 1980; 1993; Nisbett & Wilson, 2005; Willis, 2005) and this may apply to affective states and processes as well. Further, potential interviewer and mode effects may exist in questions that ask about

[7] The field of psychology deals with such problems of definition and measurement, one of the most notable of which is intelligence. Intelligence theorists (Gardner, 1999; Sternberg, Lautrey, & Lubart, 2003) have critiqued blind use of intelligence quotient (IQ) tests as sole measures of intelligence citing a logical circularity. This circularity produces statements like "What is intelligence? It's what IQ tests measure."

perceived sensitivity, just as they could on any other type of survey question. More specifically, questions asking respondents to report feelings of sensitivity may themselves be sensitive. If income is under-reported due to sensitivity, and perceived sensitivity is underreported due to sensitivity, then we don't learn much about the sensitivity of income questions by asking respondents to report how sensitive it is to discuss income. Bradburn et al. (1979) deal with this problem by asking respondents to report how sensitive or embarrassing they expect certain topics would be for other people. This is intended to deflect attention from respondents' own sensitivity, even if the construct about which they report is indeed their own sensitivity. Better measures of sensitivity and resulting affective states are likely obtained from covert measurement in the form of observations (Webb, 2000; Webb, Campbell, Schwartz, & Sechrest, 1966) including reaction times (Bargh, McKenna, & Fitzsimons, 2002; Ongena & Dijkstra, 2007) and other paralinguistic measures (Bortfeld, Leon, Bloom, Schober, & Brennan, 2001; Conrad, Schober, & Dijkstra, 2008; Schober & Bloom, 2004).

1.3.2 Income as a Type of Sensitive Question

Recognizing higher nonresponse with income data compared to other kinds of data, explanations for the higher rates have been sought in qualities of the questions. The most common tendency under a question-cause explanation of income nonresponse is to classify income questions as sensitive, based on characteristics that they share with other questions that ask about potentially-revealing personal facets like sexual behavior, drug use, or controversial opinions. This classification is justified in part by comparisons of item nonresponse rates of income questions and other sensitive questions (Tourangeau et al., 2000).

Considering the definitions of sensitivity outlined above (e.g., social desirability, intrusiveness, and risk of disclosure) income seems to meet these criteria. Reporting income can lead to socially-desirable reporting, particularly for individuals whose income is particularly low, or particularly high (Hurd, Juster, & Smith, 2003). Individuals with low income may choose to report higher incomes (over-report) or to refuse to report income because reporting their true income would involve a socially undesirable report. The reverse may happen with higher income individuals; they may under-report their income because reporting too high an income is socially undesirable as well. Income questions are sensitive under the intrusiveness explanation as well, which says that the specific true value is less important than the general cultural norm for appropriate topics of conversations with strangers (see Tourangeau & Yan, 2007 on privacy). There is a cultural norm of privacy around income and financial matters, specifically that one does not discuss income with a stranger and that it is rude for a stranger to ask about one's income. Finally, there may be issues with perceived risk of disclosure that make income sensitive for some respondents. Particularly for government surveys, some respondents may be concerned that that their income will be disclosed to other government agencies such as the Internal Revenue Service or the Social Security Administration, and this fear may be heightened for individuals on government assistance programs, those who do not file taxes, or those who intentionally misreport their income. Fear of disclosure can also extend to potential disclosure to other entities, such as marketing firms (Singer, Mathiowetz, & Couper, 1993). Further some respondents may see protection of their exact income value as a way to avoid identity theft, similar to not providing their Social

Security number, and so the psychological grounds for disclosure fear can range from fear of legal ramifications to concern about additional junk mail.

1.3.3 Complex Questions

In addition to the social psychological facets of sensitivity, survey items can vary on the cognitive effort required of a respondent to come up with an accurate answer. Increased complexity can lead to item nonresponse if the respondent recognizes the complexity and refuses to answer or says they “don’t know”. As with any kind of cognitive puzzle or problem, survey questions can tax respondents’ ability to recall facts and dates, or calculate or estimate values that satisfy the requirements of the survey question. Much research on the psychological aspects of survey responding has highlighted how cognitive difficulty can lead to error in statistics (Sirken et al., 1999; Tourangeau et al., 2000). If respondents find questions too complex to answer, they may choose to not respond at all.

The complexity of a question parallels the cognitive difficulty involved in answering survey questions that was outlined above (Tourangeau et al., 2000). Questions can vary on semantic difficulty (i.e., ability to understand what is required, including vague terminology, presupposition, or double-barreled structures (Fowler, 1995; Tourangeau et al., 2000). Questions can also ask about information that is hard to recall, either because it was never encoded, much time has passed since it occurred, or interference of other events and information has diluted the memory. Some survey questions ask respondents to make mathematical calculations or estimates (based on information that they may not have access to), such as the change in the economy over time. Any of these characteristics can vary across questions, and adds to the cognitive complexity of the questions. It is assumed that, on average, questions that are cognitively complex will lead to cognitive difficulty in answering.

1.3.4 Income as a Type of Complex Question

Income questions can be cognitively complex, as well as sensitive. Nonresponse can result from any of the stages of the cognitive response process defined by Tourangeau et al., (2000). Respondents may misunderstand the question, think it is asking about something other than total household income, think they don't have the information quested by the question, and not respond. They may understand the question, but not be able to pull the response from memory or find adding different income sources together to be too difficult, and thus not respond. They may know approximately how much they make, and be willing to share that estimate, but think that the question requires an exact figure and not respond. Different types of income nonresponse (e.g., bracketed response versus complete nonresponse, or refusals versus don't knows) may stem from different motivational forces (e.g., sensitivity and cognitive difficulty respectively), and so the distinction between cognitive complexity and sensitivity is not always clear when discussing item nonresponse.

Income questions vary widely in the type and detail of response required. A question that asks for a best estimate of household income in the past year should be simple to answer accurately, particularly if optional probes like "approximately" or "not to the penny" are included (Kormendi & Noordhoek, 1989). In that same vein, bracketed income questions also should be easier for respondents to answer than questions requiring exact income amount (Juster & Smith, 1997; Heeringa, Hill, & Howell, 1993). A series of income questions that asks for exact dollar amounts of interest income, Social Security income, and income from multiple jobs, like the income series found in the Current Population Survey (U. S. Bureau of Labor Statistics, n. d.) would be the most cognitively demanding to answer.

1.4 Question-causes, Interviewer-causes, Respondent-causes and a Psychological Model of Survey Reporting

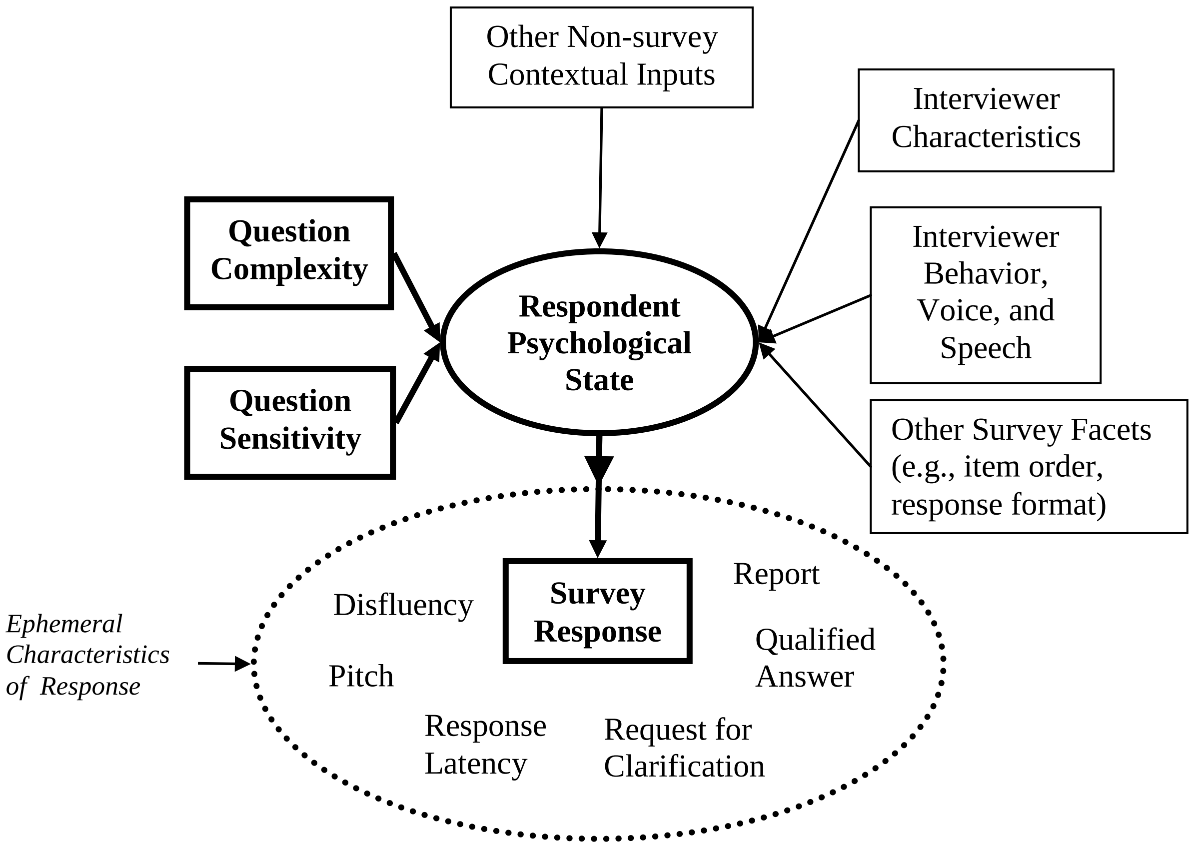
There are at least three orientations to survey response and nonresponse error that can be found in the survey methodological literature and that apply directly to the study of income nonresponse (Groves, 1989). Causes for nonresponse can be sought in characteristics of the question (including mode). Do memory demands, grammatical structure, response burden, formatting of the question, or sensitivity of the topic lead to increased nonresponse? For interviewer-administered questionnaires, causes of error can also come from characteristics and behavior of the interviewer. Do certain types of interviewers produce more nonresponse? Do certain interviewer behaviors lead to more item nonresponse? Finally (and ultimately in the causal chain), causes of error come from the respondent. A respondent's psychological state while answering the question (defined by both cognitive and affective components) may lead to item nonresponse. The respondent's psychological state will be caused in part by the mode, question, and interviewer (both whether one is present or not and if present what characteristics the interviewer possesses), as well as other personal or situational characteristics (including affective state and cognitive ability independent of the survey,).

It is important to look at respondents as potential producers of error, because all of the inputs to the survey response processes (e.g., mode, interviewer, question format) are filtered through the respondent and his or her subjective state at the moment they answer (or do not answer) each survey question. In survey methodology, we rarely have direct measures of cognitive and emotional processes (e.g., fMRI measures of brain activity or galvanic skin response), and so we are left with external, indirect indicators of affective and cognitive states. Verbal paradata offer such indicators, and are the focus of the dissertation. Psychological states can be thought of as the cause of the response error in a literal sense, and respondent paradata should covary with those states.

Figure 1 depicts different influences on a respondent's psychological state (affect and cognitive difficulty) at the time he or she responds to a specific survey question. The model assumes that all influences on the survey response are ultimately caused by the respondent's psychological state. Interviewer characteristics and behavior, topics of the survey, sponsor, mode and other survey facets only affect survey response through respondents' perceptions of these facets. The respondent's subjective psychological state directly influences the response itself (e.g., selection of category, numerical response, or verbal report), as well as the way the answer is delivered (i.e., verbal paradata). Characteristics of the answer delivery (e.g., fluency, pitch) are included in the dotted oval that surrounds the survey response. These are ephemeral aspects of the response process (i.e., potential paradata) that are lost unless captured through audio recording or real-time coding. These potential paradata are produced by the same psychological system that produces the response, and as such should covary with the response.

The primary foci of the dissertation are bolded in bold in Figure 1. The research is specifically targeted at understanding 1) the impact of question characteristics (complexity and sensitivity) on respondents' psychological states (cognitive difficulty and affect) as measured by observable indicators of those states, and 2) the effect of respondents' psychological states on their survey responses, using verbal paradata as indicators of these states. Other factors may also influence the survey response and the way it is delivered, though these are not the focus of this dissertation. Non-survey contextual inputs to a respondent's psychological state include the mood and the level of fatigue that they bring to the survey interview. Other survey facets include the mode of the interview, format of the response (open, quantitative, qualitative), topic of the question, question order, and other design features. Interviewer characteristics include interviewer sex, race, age, socioeconomic status, et cetera. Interviewer behavior, voice and speech include disfluency, pitch, speech rate, and so forth, which are many of the same conversational characteristics that are used to understand a respondent's psychological state. Interviewer behavior can also include more technical aspects of interviewing, such as whether the question was read as worded, whether response options were offered (if instructed), and whether appropriate probing was used. Each of these components work together to form a survey response system, only one small part of which will be evaluated in this dissertation.

Figure 1: Social and Psychological Inputs into Survey Response System and Associated Respondent Paradata



This dissertation draws from question-cause and respondent-cause perspectives on item nonresponse to expand and advance the discussion of income reporting. Specifically, it employs in-depth utterance-level coding and rating of verbal paradata that are hypothesized to reflect affect and cognitive states of respondents, with the goal of understanding exactly *how* income questions are answered, not just *whether* they are answered. The design also includes questions prior to the income question that have been selected based on their sensitivity and complexity characteristics. This allows for analysis of changes in respondent behavior and psychological states over questions that appear earlier in the survey and have characteristics similar to income questions.

1.5 Research Questions, Hypotheses, and Initial Data Considerations

1.5.1 Research Questions

This dissertation is guided by three primary research questions:

1) Do question sensitivity and cognitive complexity lead to more respondent paradata that reflect affective and cognitive states?
    a. If not, what latent factors seem to be influenced by question sensitivity and complexity?
2) Can a respondent's affect and cognitive state be measured through question-answering behavior, speech behavior, and voice pitch?
    a. If not, what latent factors appear to be present in respondents' answers to questions?
3) Do individual codes or latent factors derived from codes distinguish between respondents who provide a household income value, answer only in brackets, or provide no income information?
4) Can verbal paradata that occur before an income question predict the probability that income will be reported, regardless of the psychological states that they reflect.

1.5.2 Initial Data Considerations

Coding Scheme and Utterance-level Data

There are numerous ways to code interpersonal communication and numerous behaviors that can be coded. There are, obviously, the words spoken. There also are characteristics describing how those words are delivered (e.g., pacing, pausing, inflection, voice pitch). When individuals interact verbally (particularly when a visual channel is absent), there are many opportunities to interrupt each other (intentionally or not) or to talk over each other (e.g., overspeech). At even further detail, behavior like throat clearing and breathing can be coded. With so many potentially informative behaviors, actions, and characteristics to code, one decision for any communication research project is to decide which are relevant to key research questions.[8] Some research has suggested that the way a person breathes during conversation may hold semantic or pragmatic content, such as a quick breath used to take the floor from the current speaker, or a sigh suggesting difficulty in responding (Maynard, Schaeffer, Drew, Raymond, & Weinberg, 2006; Maynard & Schaeffer, 2002a; Schaeffer, 2002).

The study of communication in survey interviews has evolved over the past half-century. Perhaps the earliest approach to be developed is "behavior coding" (Cannell, Fowler, & Marquis, 1968; Fowler & Cannell, 1996). The major characteristic of behavior coding is a coding scheme that measures question asking and answering behavior (e.g., "read question as worded" or "provided codable answer"). It is also primarily used in pilot testing survey questions. Methodological studies have used behavior coding as well (Dykema, Lepkowski, & Blixt, 1997). Although behavior coding is sometimes referred to as "interaction coding" (Dykema et al., 1997) it is not always clear that "interaction" is what is being captured. Coding behavior on two parties involved in dyadic interaction and analyzing those codes does not essentially lead to an "interactional" analysis. There are qualitative (Schaeffer et al., 2004) and quantitative (Thomas & Malone, 1979) approaches that propose a more explicit focus on the interaction. Further, behavior coding

[8] See also Watt & VanLear (1996) for a contemporary review of interactional analyses in communication research.

tends to be limited to “task behaviors” (i.e., question asking and answering) in the survey process, rather than natural communication behaviors (e.g., disfluency and pitch variation) that also occur during interviewer-respondent interactions.

Another trend in research about communication in survey interviews, not completely independent of the behavior coding tradition, seeks to use natural conversation as a model for interaction in the survey interview, and to use the survey interview as a data source for testing theories about natural communication process (Kahn & Cannell, 1957; Maynard & Schaeffer, 2002a). This dissertation takes an approach to studying communication in the survey interview that captures natural communication components of interviewer-respondent interactions as well as question-answering behavior. Question asking and answering behaviors are also coded, but there is no attempt to replicate traditional behavior coding protocols (e.g., “Interviewer read question as worded”).

Chapter 2 discusses the coding scheme in more detail and Appendix C includes a copy of the specific definitions for the final codes that were used. However, it is helpful to discuss the speech and voice characteristics at a general level here to facilitate discussion of the hypotheses of this research. The following types of behavior were coded at each respondent utterance.

Question-answering Behavior:

- Respondents’ answers to survey questions and whether a qualification was present (e.g., “I guess…” or “It’s about…”)

- Requests for clarification or repeat

- Respondent uncertainty when expressed; uncertainty about the meaning of the question was distinguished from uncertainty about how to answer the question

-"Don't know" responses, distinguishing between explicit and implied don't know responses

-Refusals to answer the question, distinguishing between explicit and implied refusals

-Negative comments about the survey or the question

-Digressions from a direct question-answer path, and whether the respondent provided a codable answer in the digression

-Reports when respondents provided some information relevant to the survey question but did not directly answer the question.

Natural Communication Behavior (Speech):

-Backchanneling (i.e., active listening, such as "uh huh" in response to an interviewer's statement)

-Conversation management was a "catch-all" code that arose during the coding process and captures any communication behavior that keeps the conversation moving but that can't be coded in any other category (e.g., "Well", "How 'bout that?", "That's okay").

-Laughter

-Repairs and stammers in which the respondent either changes their statement before finishing it (a repair), or restarts a word or syllable (a stammer), with distinctions made between repairs, stammers and their co-occurrence in within an utterance

-Filler presence and duration, any word-like speech token (e.g., *um* or *uh*) that falls in the middle of an utterance

-Pause presence and duration; only within-utterance pauses were coded

-Speech rate as the number of syllables per second

-Overspeech (respondent and interviewer talking over each other)

-Number of words spoken by the respondent and by both respondent and interviewer

-Agreement or disagreement with something the respondent said, not answering a question

Natural Communication (Voice Acoustics):

-Fundamental frequency (i.e., voice pitch) median, minimum, maximum, and standard deviation at the first respondent utterance in the question

Ratings of Psychological States:

-Affect intensity on a ten-point scale (0-9), indicating how strong the affect in the respondent's voice sounded, without regard to whether it was positive or negative

-Affect valence on a three-point scale (negative, neutral, positive), indicating whether the respondent sounded like they felt good, bad, or neutral, without judging what the specific emotion was or how strongly it was felt

-Cognitive difficulty on a three point scale (no difficulty, some difficulty, high difficulty)

1.5.3 Hypotheses

This dissertation draws from diverse research literatures in survey methodology, psychology of emotions and decision making, and psycholinguistics. The hypotheses below are motivated by a general *affect/difficulty heightening* mechanism through which affect and cognitive states are increased by sensitive and cognitive complex questions, and these states increase the amount of verbal paradata produced by respondents (Bachorowski, 1999; Bortfeld et al., 2001; Conrad et al., 2008; DePaulo et al., 2003; Ehlen et al., 2007; Philippot, Feldman, & Coats, 1999; Schober & Bloom, 2004). It is also hypothesized that income nonresponse will correlate with heighted affect and cognitive difficulty indicators.

Individual measured variables (e.g., codes, speech behavior, and voice pitch) and their respective hypotheses can be organized around their expected relationship to the factors of affect (related to question sensitivity) and cognitive difficulty (related to question cognitive complexity). With respect to question characteristics, indicators that are expected to be signs of *heightened affect* are expected to be higher on questions that are high in *sensitivity* and lower on those lower in sensitivity. These indicators include implicit and explicit refusals, negative comments, backchannels, conversation management, laughter, affect intensity, negative affect valence, median voice pitch and pitch variability (e.g., range and standard deviation), rate of speech, number of utterances in answer, utterance duration, and agreement and disagreement with the interviewer. For each of these indicators, the hypothesis is that respondents will exhibit them on a higher proportion of utterances on questions that are sensitive, compared to questions that are not sensitive.

The specific measured variables hypothesized to be related to affect were chosen to represent classes of mechanisms related to sensitivity and affect. Some variables were chosen represent intentional conversational actions that reflect the sensitive or threatening nature of a question, such as refusals, backchanneling, utterance length, and number of utterances in the question. Others were selected to reflect measures of physiological arousal related to affective arousal (Bachorowski, 1999), such as laughter, voice pitch, and rate of speech. Each of these is expected to increase more or less unconsciously as sensitivity increases. Conversational and physiological variables may both increase with affective arousal. On the other hand, it could be that respondents will be *less* conversational when questions are sensitive, due either to conscious control in an effort to avoid talking about an uncomfortable question, or as a result of limited psychological resources that result from the demands of the question. Previous research does not provide clear support for reduced conversational behavior due to sensitivity, so increases due to physiological arousal will be hypothesized.

Measured variables that are expected to be signs of *cognitive difficulty* are expected to be higher on questions that are high in *cognitive complexity* compared to those low in cognitive complexity. These include answers with qualification, requests for clarification or repeat, uncertainty about the question, uncertainty about how to answer, implicit and explicit don't know, digressions, reports, repairs and stammers, cognitive difficulty ratings, within-utterance pauses (presence and duration), fillers (presence and duration), and words spoken. The hypotheses for these indicators of cognitive difficulty are that a higher proportion of respondent utterances will contain them on questions that are cognitively complex compared to those that are noncomplex.

Some of the indicators of cognitive difficulty were expected to measure expressions of uncertainty (i.e., clear, intentional conversational acts related to cognitive difficulty). These include requests for clarification or repeat, expressions of uncertainty about the question or how to answer, implicit or explicit don't knows. Others represent more subtle (perhaps unintentional) cues of cognitive difficulty, such as digressions, negative comments, conversation management, pause presence and length, filler presence and length.

With respect to question characteristics, affect, and cognitive difficulty, not all variables and hypotheses had strong motivation from theory and research. These indicators were included with the psychological state (affect or cognitive difficulty) that they seemed to best reflect. When specific empirical findings or theory did not suggest a clear hypothesis, extrapolations from theory and research were made, as well as heuristic evaluations of what these indicators might represent. Specifically, measures of the length of the exchange did not clearly fit with either affect or cognitive difficulty. To move forward with the research, an arbitrary decision was made to assign number of words spoken to cognitive difficulty, assuming that if a question is hard to answer, a respondent will either "do more things" to try to answer it, or will provide longer, more verbose answers, than if it was an easy question. Other measures of duration or speed of the exchange (e.g., number of utterances, speech rate) were assigned to affect. Speech rate was thought to measure fast, nervous talking, while number of utterances was thought to measure exchanges where respondents were evasive due to discomfort with the question.

There are also hypotheses about the relationship between income nonresponse and measured variables on questions earlier in the questionnaire. The indicators hypothesized to be related to complete income nonresponse are primarily those that indicate affect, and it is predicted that income nonresponse will be due to affective reasons. Variables hypothesized to be positively associated with income nonresponse are implicit and explicit refusals, negative comments, negative affect valence, affect intensity, median voice pitch and voice pitch variability (e.g., range and standard deviation). The hypotheses for these indicators is that they will be more prevalent before the income question in respondents who eventually do not provide any income data (e.g., income nonrespondents) compared to bracketed respondents and dollar amount respondents.

Cognitive difficulty is hypothesized to lead to the use bracketing, and so the indicators hypothesized to be positively associated with bracketing are requests for clarification and repeat, expressions of uncertainty about the question or how to answer, digressions, backchanneling, conversation management, reports, repairs, stammers, cognitive difficulty ratings, within utterance pause presence and length, and filler presence and length. All these indicators are hypothesized to be more prevalent before the income items in respondents who eventually report income with a bracketed response, compared to nonrespondents and dollar amount respondents.

A single hypothesized mechanism for income dollar amount response is not clear from the literature, hence the motivation for this research. Income nonresponse can be due to either discomfort with the question (e.g., sensitivity and affect) or problems coming to an answer (e.g., cognitive complexity and difficulty). It is expected that several indicators of affect and cognitive difficulty will predict income nonresponse, and that those who provide income will have an absence of most difficulty and affect indicators. The variables hypothesized to be related to income dollar amount reports are primarily affective and conversational, including backchanneling, conversation management, laughter, affect valence. To summarize, where prediction of income nonresponse and bracketed response can be made, it is anticipated that measures of affect will predict income nonresponse, and measures of cognitive difficulty will predict bracketed response, as brackets are thought to aid respondents who have a difficult time arriving at a specific income value.

The remaining chapters present the data development and statistical results relevant to these hypotheses. Chapter 2 reviews the coding scheme and data development. Chapter 3 looks at differences in rates of verbal paradata related to question sensitivity and complexity. Chapter 4 explores the relationship between verbal paradata and income nonresponse. Chapter 5 concludes with a synthesis of the findings and extensions into future areas of research on income nonresponse, verbal paradata, and psychological processes behind reporting sensitive and complex information.

Chapter 2

General Method

The raw data analyzed in this dissertation, come from audio recordings of selected questions from interviews conducted by the Reuters/University of Michigan Surveys of Consumers (SCA). The income question in this survey asks respondents to report their household income in the past calendar year (e.g., income in 2009). Their response can be an exact or estimated dollar amount, a bracketed dollar amount, or no response at all. A sample was drawn that includes roughly equal numbers of respondents who answered income in each of these three categories. Only random digit dial (RDD) cases were selected. Measures of affect and cognitive difficulty before the income question come from audio recordings of interviewer-respondent interactions on four questions that were selected for sensitivity and complexity as judged by reviewers. The four questions before income create four conditions, sensitive and complex, sensitive and noncomplex, nonsensitive and complex, and nonsensitive and noncomplex, that can be described by two within-subjects factors (sensitivity and complexity) with two levels each (presence and absence).

The SCA records all interviews unless the respondent explicitly says the do not want to be recorded. Thus audio data are available for all respondents. Audio recordings of the interaction between the interviewer and respondent were transcribed verbatim for each selected question. After transcription, they were coded for question-asking and question-answering behavior, speech, and perceived psychological states of respondents and interviewers. Voice pitch (fundamental frequency) was measured electronically from the recordings. Question-level measures were then calculated (e.g., the proportion of utterances on which the respondent laughed) and these measures are the data analyzed in the rest of the dissertation. This chapter will describe the selection, coding/rating, and pitch measurement protocols and procedures, and the resulting data structure. All of the data analyses reported in chapters 3 and 4 will be based on the final sample, but this chapter documents the coding scheme development, testing, and training, which includes a practice sample. Transcription will be discussed first, then coding, then sample selection.

2.1 Transcription and Coding Scheme

2.1.1 Transcription

Recordings were first transcribed verbatim for reference when coding. Transcription primarily included text of spoken words, with minimal markup for things like pauses, and interruptions. Interviewer and respondent speech were marked with the letters “I” and “R” respectively. In the practice phase, transcripts were created demarcating the speaker at the turn level. All interviewer speech beginning the question exchange would receive an “I”. When the respondent began speaking, their speech would be noted with an “R”, until the interviewer began speaking again. The following is an example of a transcript demarcation by conversational turn.

I: What was your family’s total income in 2007?

R: Well, that’s a complicated question for us. Can you re-read the question?

I: Sure. What was your family’s total income in 2007?

Conversational turn taking is a major structural component of interpersonal communication. A turn defines the time during which one conversational partner holds the floor (i.e., is speaking and the second partner is not speaking). However, multiple behaviors or speech acts may need to be identified within a single conversational turn (e.g., expression of confusion and asking of a question). To be able to isolate individual behaviors of interest, and be able to apply more discrete codes to more precise speech segments, utterances can be demarcated. Each conversational turn can have multiple meaningful utterances, where meaning is defined by research questions and a particular coding scheme. In the survey context, for example, a respondent might express difficulty answering and ask for a repeat of the question both in one turn. In the dissertation, these conversational components are referred to as “utterances” (Fromkin, 1973; Klatt & Klatt, 1990; Schober & Bloom, 2004). The example above is re-transcribed so that utterances are noted individually.

I: What was your family’s total income in 2007?

R: Well, that’s a complicated question for us.

R: Can you re-read the question?

I: Sure.

I: What was your family’s total income in 2007?

Based on the complication of developing a simple coding scheme that captured all relevant behavior, the decision was made to transcribe and code the real sample at the utterance level (rather than the turn level), so that one code could be applied to each meaningful utterance. In addition to interviewer and respondent behavior, speech characteristics, pauses, affect, and voice pitch were measured at the utterance level, where a single conversational turn can contain multiple utterances. During the practice phase, it quickly became apparent that there would be many turns that contained multiple actions of interest given our coding scheme. In a turn-level coding scheme, the only way to deal with multiple behaviors in one turn is to create a code that captures multiple actions of interest whose (e.g., “X and Y occur”). While this is possible it is not ideal for two reasons. First it loses the temporal sequence of events X and Y, unless separate codes are created for each order (e.g., “X then Y”, and “Y then X”). If three behaviors are present in the turn, the complexity of the codes quickly compounds (e.g., “X and Y”, “X and Z”, “Y and Z”, etc). Second, code purity (i.e., one code for one behavior or action) is reduced and analytic complexity (i.e., need to recode to isolate individual variables) is added when codes contain multiple behaviors of interest. It becomes more difficult to separate each individual behavior and analyze it separately. Further, the range of possible codes increases exponentially. Consider the codes needed if three variables can co-occur (e.g., A only, B only, C only, A and B not C, A and C not B, B and C not A, A B and C). Utterance-level transcription and coding displayed here were used in the full sample, but only turns were demarcated in the practice coding.

Minimal markups were applied to the transcripts as a way of coding some of the speech and communication phenomena of interest. Overspeech, where the two speakers talk over each other (i.e., their speech occupies the same temporal location) was transcribed by putting asterisks around the specific words or syllables that overlapped. Interruptions (either self or other) were transcribed with a hyphen (e.g., “-“). This mark-up was used when the next speaker’s utterance started at a point that seems to be in the middle of the current speaker's utterance, but no overspeech was present (i.e., the current and next speaker don't literally talk over each other). It was also used for self-interruptions that accompany repairs and stammers. Transcripts were also marked for pauses of one second or more. A period surrounded by spaces on either side (e.g., " . ") denoted a pause. Lengthened sound was marked by inserting a colon within the word at the point the sound was lengthened (e.g., ":"). Rising intonation was noted by the use of a question mark at the end of a sentence (e.g., "?"), whether the sentence was declarative or interrogative. The initial transcription protocol used in the practice phase is in Appendix A.

Each of the 26 training interviews were transcribed and verified by a second transcriber who listened to and corrected the transcript according to the transcription protocol if necessary. Disagreements were reconciled when needed, and this verified transcript was the one that was eventually coded.

2.2 Coders, Coding Process, and Reliability

Ten students (9 undergraduates and one recent graduate at the University of Michigan) worked on the development and implementation of the coding scheme. Students were mostly juniors and seniors, most with some previous research experience. Two were first-year students in the University's Undergraduate Research Opportunities Program (UROP). Most students were social science majors (a majority in psychological sciences), but others were studying language, health sciences, and business.

Coding commenced based on a coding scheme that was developed before coders were involved and without substantial testing. Neither the original transcription instructions nor coding scheme had any input from the coders. The goals of the practice phase were to test and practice these protocols, as well as add or remove protocols as seemed appropriate. The original coding scheme can be seen in Appendix B.

All items were coded by two coders so reliability of the coding scheme could be measured. Coding was done in pairs, so each coding pair coded about 1/3 of the practice sample. Reliability varied across codes, as might be expected. The overall reliability of the coding scheme, calculated as a weighted Cohen's Kappa[9], on all codes across all utterances (e.g., reliability of event codes) was .603, .51, and .529 for each of the three pairs of coders. Reliability for specific code variables (e.g., actor, behavior, anxiety, and cognitive complexity) varied between pairs, and ranges are shown in Table 1. Codes of observable behavior, such as interviewer and respondent behavior, or presence and absence of specific speech acts were relatively reliable (between .411 and 1.0 removing outlier pairs). Respondent negative comments, for example, had reliabilities ranging from .787 and .922. Reports had reliability ranging from .499 to .885. Ratings on less objective characteristics (e.g., ratings of anxiety or cognitive difficulty present in speech and professionalism of the interviewer) were less reliable (from .222 to .587 across variables and coder pairs). Debriefing of coders offered some insight into how they used the coding scales and definitions for the subjective judgments, and those comments were used to modify the coder training before the full sample of 185 SCA cases were coded. It was expected that more training on these variables, and possible re-definitions to match coders' intuitive impressions would lead to more reliable codes. The lower reliability on subjective judgments of affect and cognitive difficulty is likely due in part to the fact that

[9] The weighted kappa in Sequence Viewer gives more weight to utterances for which more codes (e.g., actor, behavior, laughter) are identical between the two coders. Weight in the reliability calculation is proportional to the number of codes (columns) that are the same between two coders.

they are true judgments, and have less explicit definitions than other codes. What sounds like anxiety to one coder might not sound like anxiety to another. For question-answering behaviors and speech acts on the other hand, coders were looking for specific observable behavior that fit a fairly strict definition. For observable behavior, a judgment must be made about whether the observed behavior fits the definition, but the amount of judgment is likely less than is involved in judging anxiety, difficulty, and professionalism. Further the subjective judgments were applied on a scale, rather than by presence or absence. The use of a scale alone allows for more variability in response.

Table 1: Reliabilities of Individual Code Variables in Practice Sample

Type of Code/Judgment	**Code Variable**	**Reliability Range for Coding Pairs**
Objective behavior or component of interaction	Actor	.98-1.0
	Behavior	.60-.76
	Respondent Comment	.79-.92
	Pause	.03-.41
	Report	.50-.89
	Repair	.51-.59
Subjective judgment	Anxiety	.22-.31
	Cognitive Difficulty	.34-.59
	Interviewer Professionalism	.31-.40

Table 1 provides strong evidence for re-evaluating the coding method for pauses. The most reliable pairs had reasonable reliability, but the wide range of kappa coefficients, including one pair that had a kappa less than 0.1 suggests difficulty with the coding scheme. The code for pauses seemed to produce major problems, as measured by kappa (.030 for one pair), while other objective codes showed moderate or high reliability. Codes for subjective judgments were less reliabile on average than those for objective behaviors. Looking at the kappa coefficients of subjective judgments for the most reliable pairs (.309, .587, and .4) suggests that moderate to high reliability is possible. The reliability results for anxiety and cognitive difficulty codes prompted a retraining and restructuring of the codes before the full sample was coded. Interviewer professionalism, though moderately reliable across pairs was dropped because no clear hypotheses were present.

Debriefings with coders and further review of the literature made it evident that there were some oversights in the original coding scheme. One goal of the practice coding phase was to find such problems, and refine and change the coding scheme accordingly. We made one round of revisions to the coding scheme and tested that on the practice data. Those revisions can be seen in the final coding scheme in Appendix C, and include additional codes for interviewer and respondent behavior that were not in the practice coding scheme but came up frequently enough to warrant their own code (e.g., respondent ask for clarification, interviewer says "thank you"). We also created an utterance-level code for laughter which had previously been assigned a question-level, and coded pauses and fillers in a Sequence Viewer function called "Time Keys" that will be discussed later.

Notable results and modifications that came out of the practice coding are as follows:

1) Settling on a reduced transcription protocol that captured speaker, interruptions and stammers, and overspeech. Transcription marks for intonation and typing were dropped from transcription due to the time it took to code these additional components, and the fact that they were not central to any of the hypotheses in the dissertation.

2) The transcription and coding schemes were changed from turn-level to utterance-level.
3) Utterance level codes were added including:
    a. Laughter was changed from a sequence level (question level) code to an utterance level code.
    b. Conversation management was added as a code that could catch all statements that moved the conversation forward, but did not have a clear place in the more substantive codes.
    c. The anxiety code was changed to an affect code and split into two items: affect intensity and affect valence. This resolved some of the trouble coders expressed in trying to identify anxiety, and reflects the literature on coding affect in speech (Bachorowski, 1999).
    d. The cognitive difficulty code was changed from a 5-point scale to a 3-point scale, which helped coders assign codes. Coders had a hard time distinguishing moderate levels of cognitive difficulty. Also, retraining on this variable showed that some coders were imposing a normal distribution on the 5-point range, calling 3 an "average" level of difficulty. Using a three point scale and labeling the points "no difficulty", "some difficulty", and "high difficulty" seemed to reduce or eliminate this problem.

Two major misunderstandings about the coding scheme were discovered during weekly coder meetings and retraining. These misunderstandings were clarified, and remained a regular topic of our weekly meetings. First, some coders wanted to interpret speakers' intent beyond how the spoken words sounded. For example, one coder once noted, “That interviewer SOUNDS happy, but I think she’s actually irritated at the respondent.” While it was expected that coders would be listening for qualities of voice, it was not expected that coders would be “reading into” the psychological state of the respondents beyond general ratings of affect and difficulty. The coding instructions stated to code affect and difficulty from the voice alone, and this was clarified in retraining and throughout the coding process.

Second, coders sometimes took the perspective of the interviewer or respondent and coded based on what they would feel if they were interacting with the other party. Coder comments like “I would be really irritated if that was MY interviewer” or “Something about that interviewer just bugs me” were common during our debriefing and retraining. Coders’ perceptions about individual cases didn’t always align with each other, despite the goal of this coding scheme to be as objective and reliable as possible. Such subjective differences between coders in their perception of and inference about the observed behavior are what make obtaining a reliable coding scheme such a difficult task. We came up with two coding rules of thumb to which the coders could agree and about which they were reminded regularly; 1) code what you hear, not what you infer about the psychological state of the speaker (i.e., do they “sound happy”, not “they sound happy but it’s a fake happy”), and 2) do not place yourself in the role of the interviewer or respondent (i.e., step back from the exchange and code what you hear in each speaker’s voice, not what you would be feeling if you were talking to that particular interviewer or respondent).

Coding fillers and pauses was somewhat more straightforward. Fillers and pauses are both points in speech (often within an utterance or turn) where words are not being spoken, but the floor is being held (i.e., the next speaker has not begun to speak). They are both considered types of speech disfluency. In this study, pauses are completely soundless space between speech fragments, whether within utterances or between utterances. They may include only breathing but no other speech-like sound like *um* and *uh*. Fillers were defined as any sound, such as *um*, *uh*, *er*, *ah*, *eh*, that was not a backchannel, but seemed to be placed within the utterance as a "placeholder". Coders were trained to mark fillers objectively, without consideration of their purpose within conversation, other than to distinguish something like *ah*, which could be a backchannel, from *ah* as filler. Other than the instruction not to consider breathing, sighing, etc (even if they thought it was serving the same purpose as a filler), coders were given significant leeway about what to code as a filler, realizing that there are more speech tokens that can be used as fillers than just *um* or *uh*. Appendix D shows the protocol for coding fillers and pauses in Sequence Viewer.

While question-answering behavior, speech, affect, and cognitive difficulty were coded or rated discretely at each utterance (i.e., utterance received one and only one code on each dimension), the coding protocol was different for pauses. The goal was to have a coding system that 1) distinguishes between empty pauses (referred to as pauses) and filled pauses (referred to as fillers), 2) allows within-utterance pauses to be distinguished from between-utterance pauses, 3) allows for calculation of presence and duration of pauses and fillers, and 4) allows pauses and fillers to be attached to specific utterances, and thus specific event codes (e.g., attach filled pauses to refusals at the same utterance). The Time Keys function in Sequence Viewer accomplishes this, and is implemented independently of the coding of utterance-level event variables that has been discussed thus far. In the end, the data are tied together by time markers that are applied to the audio recordings.

Using Time Keys, a coder marks the beginning of the pause or filler they wish to code. They then put a second mark at the end of the pause or filler. The onset mark, offset mark, and all space in between are then denoted as a filler. Statistics like total count of fillers, total duration in filler, proportion of utterances with a filler, proportion of fillers per utterance, proportion of time spent in fillers, can then be calculated. If time markings are attached to the events also, time key presence and duration can be attached to specific events (utterances), and thus event codes given to those events.

In some literature, a rule of thumb for marking pauses of 1 second or more is used (Schober & Bloom, 2004). In this project, the marking of within-utterance pauses was left more in the discretion of the coder, with the one instruction being that anything 1 second or more should be marked with certainty. Pauses below 1 second and length were marked at the coder's discretion, with instruction to mark more rather than fewer pauses.

Time markings, fillers, and pauses (e.g., all Sequence Viewer components that are not code variables) were implemented under a truncated coding scheme that did not include all utterances within each question, in order to save coding time and also be likely to capture the utterance in which the question was answered. Coders were instructed to put timestamps on utterance beginnings and endings (defined by words only), and mark fillers and pauses through the first four utterances (interviewer and respondent) that are not overspeech utterances. It is expected that any reaction to question sensitivity or complexity that is manifest in speech and voice would happen either immediately after hearing the question, or during the next one or two turns, in which the respondent answers, refuses, expresses confusion about the answer, et cetera. This reduced coding plan was implemented to maximize the possibility of capturing this variability while reducing coding time relative to marking and coding all utterances.

Coders were not completely blind to the income nonresponse status of respondents. They knew the case identification system (i.e., the ID number that identified each case as an income nonrespondent, bracketed respondent, or dollar amount respondent). They also were able to hear the question on which respondents were asked income, and so could have deduced which respondents were nonrespondents, bracketed respondents, and dollar amount respondents. In this study, one risk to the objectivity of the coding scheme was related to coders' knowledge of the "history" and nonresponse status of each case (i.e., the repeated measures for each subject). It was feared that if coders knew too much about each case, they might code later behaviors of the same respondent by thinking of earlier ones (e.g., this person sounds *more* confused now than earlier), thus introducing dependence of the codes due to coder within a case across the repeated measures. To help minimize any influence of earlier questions on later questions, individual questions were coded in such a way that each coder coded a set of recordings of one question (e.g., Question 1 only) and different respondents, rather than coding all the questions for each respondent (e.g., questions 1-5 for respondent 101) sequentially. Other than this stipulation cases were assigned to coders in a practical way, based on which cases were prepared and which coders were available to work on them.

Coders were also instructed to code question-answer behaviors and speech facets (repairs, reports) in one pass, and the more subjective facets of the interaction (affect and cognitive difficulty) in a separate pass. It was found from discussions with coders that this allowed them to listen more carefully to speech in one pass, and “tone of voice” in another, hopefully insuring the accuracy and purity of the codes.

Finally, coders were kept completely blind to the hypotheses of the study. They were never told what the project was looking for, or what previous research suggested. We allowed discussion of folk-theories during our meetings, which gave coders a chance to develop their own insights about what was happening in the interactions. They seemed to need little encouragement to do this. Although folk theories were developed, for any coder who thought a speech behavior might be related to nonresponse, there was another who thought it wasn’t. Discussion in our weekly meetings always came back to the fact that (at least for the time being) we didn’t know what related to what, and we needed to code the cases objectively based on the coding scheme.

Reliability coding of the real sample was conducted by a coder who did not conduct any of the original data coding. It was thought that this might help establish the validity of the coding scheme. Two coders who were both part of the project all year might produce high reliability simply because they have each listened to similar cases, each been to the same meetings, and have each had the same amount of time to practice, implement, and think about the coding scheme. They may also have picked up subtle idiosyncrasies that were not clearly reflected in the coding protocol, but arose out of the project meetings and talking to each other between meetings. Worse, they may have developed their own coding conventions that actually contradict the coding scheme. Coders relying on the same rules will lead to high reliability, even if those rules do not reflect the coding instructions. That is, coded data can be reliable without being accurate. However, if a new coder can produce reliable data using only the approved coding scheme (e.g., not *ad hoc* conventions developed by coders themselves), the evidence for a good coding scheme is stronger.

A fraction of the original questions were randomly selected so that each original coder was represented in the reliability coding. Cases were selected from each coder for a total of 144 cases to code for reliability. All original codes (including speaker designation) were removed from the Sequence Viewer file, so that the reliability coder had no knowledge of how the cases were originally coded, other than the original transcription and utterance demarcation. Table 2 shows the reliabilities for each column in the Sequence Viewer file. Reliability coding was not conducted for marking of pauses and fillers, as they were marked in a feature of Sequence Viewer that does not make reliability calculation simple. Also, it is reasonable to think that pauses and fillers, as relatively objective phenomena, should be coded consistently across coders. Pitch measures were not reliability coded because they were extracted mechanically using acoustic processing software, and thus are not prone to coding error.

Table 2: Reliabilities of Individual Code Variables in Full Sample

Type of Code/Judgment	Code Variable	Kappa	Percent Agreement
Objective behavior or component of interaction	Actor	.99	99.3
	Interviewer Behavior	.90	91.9
	Respondent Behavior	.77	85.3
	Laughter	.61	98.2
	Report (respondent only)	.83	91.2
	Repair and Stammer	.65	93.6
Subjective judgment	Affect Intensity	.03	49.6
	Affect Valence	.28	67.8
	Cognitive Difficulty	.21	86.0

Kappa and percent agreement often tell different stories about intercoder reliability, depending on the specific distributions of the variables. Visual exploration of the cross tabulation of the two coders, in conjunction with agreement statistics, can provides more information about the agreement of particular categories in the coding scheme, and can be done by reviewing Appendix F. For example, looking at the “kappa per code” column for respondent behavior (Appendix F.3) we can see that the kappa of .772 for the entire variable varies widely by particular category (code). Coders coded answering the question (with and without qualification), don’t know answers, and digressions more reliably than they coded digressions with a codable answer, and backchannels. Looking at reliabilities for coder pairs, and the variability across pairs of coders gives some insight into the strengths and weaknesses of the coded data for individual codes.

In addition to code-by-code reliabilities, Sequence Viewer calculates overall reliability, considering each unique string of codes (e.g., 9 variables coded per utterance) as a code, calculating a traditional kappa, and weighting the kappa by the proportion of columns that are similar. For example, in the weighted kappa, codes are considered "more similar" if 8 of 9 codes are the same, and less similar if 1 of 9 are similar. The overall weighted kappa for the entire coding scheme is .53, which can be considered moderate. Reliabilities were somewhat improved over practice coding, though not as much as was hoped for the subjective judgments of affect and cognitive difficulty. Still, most of the coders are within moderate to high (i.e., acceptable ranges).

2.2.1 Pitch Extraction

The Praat software (Boersma & Wenak, 2009) is used to objectively measure voice pitch in selected utterances of respondents. Pitch measurement is one type of measurement in a broader class of analyses referred to as acoustic analysis. Pitch is the only acoustic voice characteristic that will be analyzed in this dissertation. The terms pitch and f_0 (fundamental frequency) will be used interchangeably unless the context requires one or the other. Fundamental frequency (measured in Hz) is the acoustic component of voice that is measured in Praat when "pitch measures" are discussed. It is the acoustic term for the component of voice that humans hear is pitch.

Praat provides a measure of pitch that is based on an analysis of the sound wave present in a sound file. Scripts can also be written to extract pitch measures in the form of distributional statistics from batches of audio files. Praat requires a few parameters to be set before pitch can be measured, including pitch ceiling and floor (defining the upper and lower limits beyond which f_0 measures won't be recognized) and voicing threshold (defining how loud a sound needs to be before an f_0 value is calculated).

The actual analysis of f_0 and production of distributional statistics from a sound wave is a fairly complex statistical process, accomplished by the Praat software. It essentially involves autocorrelation and cross correlation procedures that are applied to the frequency signal produced by the sound sample (Boersma, 1993). Cross-correlation was used in this analysis, compared to other autocorrelation settings available in the software. This is a recommended autocorrelation setting for the analysis of f_0 in Praat and simply warrants documentation.

Not all utterances were marked and analyzed in Praat. To match the protocol described earlier for fillers and pauses, only the first 4 non-overlapping respondent utterances were marked in Praat. Thus, utterances were marked in the Praat program to be identical to the utterance markings in Sequence Viewer. Thus, the resulting data were completely compatible with each other, and could be merged to an external data set. Utterance 1 in Sequence Viewer was also utterance 1 in Praat, and so on. Between-utterance pauses were marked in Praat as well, so that only spoken words were highlighted for pitch extraction.

Using some previously developed Praat scripts (Benki, 2005a, Benki, 2005b), a number of acoustic variables were extracted from the sound files, including median f_0, 5^{th} and 95^{th} percentile of f_0, f_0 range (calculated from 5^{th} and 95^{th} percentile), standard deviation of f_0, and f_0 over the last 50 ms of voicing. Some non-acoustic speech variables are calculated as well, first utterance duration, rate of speech at first utterance, and number of respondent and interviewer turns in the question. All acoustic and non acoustic variables derived from Praat are calculated only on the first respondent utterance in the question.

Although mechanical pitch analysis is preferable to human coding of pitch because it removes measurement error due to human judgment, there are still some aspects of the analysis that are prone to error. Specifically, pitch doubling and halving are phenomena that can occur in the mechanical analysis of pitch. Pitch analysis is also sensitive to the quality of the recording, which can be influenced by the medium onto which the recording was made, the recording device (microphone), and background noise. Background noise in particular needs particular attention in pitch analysis. Imagine that a respondent has a pitch ranging between X and Y. In the middle of their speech a child, who is not part of the conversation, starts singing a song. If the Praat software picks up both the speaker and the child, the pitch measure can be distorted, reflecting the higher pitch of the child. Similarly, birds, dogs, cats, squeaky doors, knocking on tables, clacking keyboards, and the whir of a tape recorder can perturb the audio recording and extraction of pitch. Higher pitch also occurs in natural, speech on certain consonant sounds (e.g., "s"). Further, pitch doubling can occur as a result of the measurement of f_0 within Praat, due to segments of voice that fall outside the settings of the program. Pitch doubling on cases in the first marking was checked and fixed where found. A second pitch marking was done, in which the first respondent utterance on which the respondent provides an answer, attempts to answer, or refuses the question was also done, though those markings have not been analyzed. The f_0 of the literal first respondent utterance is analyzed in the dissertation.

2.3 Data Source: Reuters/University of Michigan Surveys of Consumers

The SCA conducts about 500 monthly household telephone interviews with respondents selected through a random digit dial (RDD) sampling method. About three-fifths of these are new RDD cases each month and two-fifths are recontacted cases who were first interviewed six months prior. For the dissertation, RDD interviews from October 2007 through October 2008 serve as the sampling frame. The selection criterion for cases was income item nonresponse status, where respondents could either refuse to report income data (income nonrespondent), report it within a range of values (bracket respondent), or report a dollar amount (dollar amount respondents).

Several qualities of this data source make it ideal for studying the impact of speech and voice on income item nonresponse. The first is its size, with (about 300 new RDD interviews each month across about 25 interviewers). The second is its wide range of income item nonresponse, from 7 to 33% for refusals to the open-ended income question and 3 to 20% for final refusals after being offered bracketed income ranges in any given month (see Yan et al., 2006). The third is the digital format of audio recording; all recording has been done digitally so no digitization is needed. Further, with its focus on income and financial topics, and the economy more broadly, this study provides questions that range in sensitivity (e.g., questions related to income, income changes) and complexity (e.g., questions asking for knowledge and prediction about economic conditions). Finally, as part of an active field survey, the data source has more natural external validity than comparable laboratory research. However, because no validation data are available, it is not possible to examine the relationship between speech and voice qualities on response accuracy.

2.3.1 Case Selection

The Survey of Consumers provided 30 cases from the October 2007 survey that served as the data for development and testing of the coding and transcription schemes, and training of coders. Four recordings were corrupted or had other errors and could not be used. The final training sample consisted of 26 cases, 8 open-ended income respondents, 8 bracketed respondents, and 10 income nonrespondents. No analyses, other than reliability calculations, were conducted on the practice questions and cases. They were used solely for practice and refinement of the transcription and coding scheme.

The SCA recordings used for the dissertation research come from one year of Survey of Consumers monthly interviews (October 2007 through September 2008). The selection was limited to random-digit dial cases, excluding recontact cases who were first interviewed 6 months previous. This excludes those who have prior experience with the SCA and thus might be more comfortable and accustomed to the survey interview, thus being more likely to report income and showing less discomfort or difficulty in their speech, voices, and question-answering behavior. Also, since most surveys are one-time interviews, the external validity of the findings benefits form only using RDD cases.

Cases were selected based on whether their income was reported in an open-ended format, in a bracket format, or not reported at all based on the SCA data files. The proposed sample size (200) was divided roughly into three groups of 70 cases each, for a total planned sample size or 210. Upon reviewing SCA cases from October 2007 through September 2008, it was found that only 63 bracketed respondents were available for the whole selection period (of 3584) cases total. While this seems like a very small number, it was noticed anecdotally that some respondents began the bracket series, and even answered one or two brackets before provided a dollar amount for their income response. For respondents like this, it seems that the brackets helped them answer income, and avoided nonresponse, even if their response in the end was a dollar amount and not a bracket. These respondents were considered dollar amount respondents in this study.

For the other two categories, dollar amount respondents and income nonrespondents, 70 cases were randomly selected from total available cases of 3184 and

337 respectively. Unfortunately, not all of these cases were usable, and re-requests had to be made for additional cases to retain the original sample sizes.

Table 3: Selection Rates of Each Income Nonrespondent Type

	Income Nonrespondent	**Bracketed Respondent**	**Dollar Amount Respondent**
Number selected	70	63	70
Total	337	63	3184
Percent selected	20.77%	100%	2.20%

The final data set used for analysis removed cases that didn't have a recording for each of the five questions and cases that had other problems with the recording. It includes 185 respondents.

2.3.2 Item Selection

Four items plus the income item were used in the practice coding. The four occurred prior to income in the SCA interview, and were selected for sensitivity and complexity based on the investigator's judgment. Four different questions, selected for the same characteristics, were used in the full sample coding. The household income question remained the same between the practice and full sample coding phases. The four practice items are below in the order they appear in the survey (response options are shown in ALL CAPS.

Practice Question 1: Sensitive but Noncomplex

"We are interested in how people are getting along financially these days. Would you say that you (and your family living there) are better off or worse off financially than you were a year ago?"

BETTER OFF, SAME, WORSE OFF, DK, NA

”Why do you say that? (PROBE: Are there any other reasons?)”

OPEN-ENDED RESPONSE

<u>Practice Question 2: Nonsensitive and Noncomplex</u>

“During the last few months, have you heard of any favorable or unfavorable changes in business conditions?

YES/NO

“What did you hear? (Have you heard of any other favorable or unfavorable changes in business conditions?)”

OPEN-ENDED RESPONSE

<u>Practice Question 3: Complex but Nonsensitive</u>

“During the next 12 months, do you think that prices in general will go up, go down, or stay where they are now?”

PROBE “SAME” RESPONSE: “Do you mean that prices will go up at the same rate as now, or that prices in general will not go up during the next 12 months?”

GO UP, STAY THESE SAME, GO DOWN, DK, NA

“By about what percent do you expect prices to go (up/down) on average during the next 12 months?”

0-94%, 95% OR MORE, DK, NA, INAP

Practice Question 4: Complex and Sensitive

“During the next 12 months, do you expect your (family) income to be higher or lower than during the past year?”

HIGHER, SAME, LOWER, DK, NA, INAP

“By about what percent do you expect your (family) income to (increase/decrease) during the next 12 months?”

0-94%, 95% OR MORE, DK, NA, INAP

Household Income Item

“To get a picture of people's financial situation we need to know the general range of income of all people we interview. Now, thinking about (your/your family's) total income from all sources (including your job), how much did (you/your family) receive in the previous year?”

OPEN-ENDED RESPONSE IN DOLLARS

IF REFUSAL OR DON’T KNOW: BRACKETED FOLLOW-UP FOR NONRESPONDERS.

“Is your household income above $50,000?”

IF YES: “Is it above $60,000?”

IF NO: “Is it above $10,000?”

RESPONDENT CONFIRMS INCOME IS ABOVE A CERTAIN DOLLAR AMOUNT; BRACKETS CONTINUE UNTIL RESPONDENT REFUSES TO ANSWER FURTHER

For the items analyzed in the dissertation, a selection plan was used that attempts to maximize the variability in respondent voice and speech prior to the income question and does so in a way consistent with the hypotheses about the relationships between question complexity and sensitivity and latent psychological states. Questions were selected that would have the most potential to increase indicators of cognitive difficulty and potential to increase indicators of affect. Four SCA questions were selected based on expert and novice ratings; one that was cognitively complex and sensitive, one that was cognitively complex but not sensitive, one that was sensitive, but not cognitively complex, and one that was neither cognitively complex nor sensitive.

For the full sample, question selection was made by asking a set of novice and expert coders to rate the perceived complexity and sensitivity of all the questions in the SCA interview schedule prior to the income question. Question sensitivity and complexity were both operationalized on a 10 point (0-9) scale where 0 indicated the absence of sensitivity and complexity. An independent samples t-test was run for each question comparing novice and expert responses. Ratings were statistically similar except for a question asking about whether it was good time to buy a house, which was rated as more cognitively complex by novices (all of whom were younger than the experts). Averages per item were then calculated across all rates and sorted to identify the items with the highest sensitivity (lowest complexity), highest complexity (lowest sensitivity), highest sensitivity and complexity, and lowest sensitivity and complexity ratings. These ratings then determined the selection of the sensitive, complex, sensitive and complex, and neither sensitive nor complex items. The questions are listed below in the order they appear in the survey. The income question is the same as that used in the practice coding and appears as the fifth question in the series of five.

Question 1: Complex, Nonsensitive

"What about the outlook for prices over the next 5 to 10 years? Do you think prices will be higher, about the same, or lower, 5 to 10 years from now?"

HIGHER, SAME, LOWER

IF "SAME": "Do you mean that prices will go up at the same rate as now, or that prices in general will not go up during the next 5 to 10 years?"

SAME RATE, PRICES WON'T GO UP

"By about what percent per year do you expect prices to go (up/down) on the average, during the next 5 to 10 years? (How many cents on the dollar per year do you expect prices to go (up/down) on the average, during the next 5 to 10 years?)"

PERCENT OR CENTS ON THE DOLLAR

Question 2: Sensitive, Complex

"During the next year or two, do you expect that your (family) income will go up more than prices will go up, about the same, or less than prices will go up?"

MORE THAN PRICES, SAME, LESS THAN PRICES

<u>Question 3: Sensitive, Noncomplex</u>

"During the next 12 months, do you expect your (family) income to be higher or lower than during the past year?"

HIGHER, LOWER

<u>Question 4: Nonsensitive, Noncomplex</u>

"Speaking now of the automobile market - do you think the next 12 months or so will be a good time or a bad time to buy a vehicle, such as a car, pickup, van, or sport utility vehicle?"

GOOD TIME, BAD TIME

"Why do you say so? (Are there any other reasons?)"

OPEN-ENDED RESPONSE

2.4 Resulting Data Set

The resulting coded data set consists of 194 respondents with at least one question. A few respondents were not recorded on all of the questions or had recordings that were corrupted and could not be used, so the resulting analyzed data set includes 947 questions. In terms of utterances, there were 9090 interviewer and respondent utterances (3799 of which were respondent utterances, 5275 were interviewer utterances and 16 belonged to a third party). Some respondents were missing one or more of their repeated measures (e.g., questions), and thus the total number of respondents that had five repeated measures was 185 (925 repeated measures, 5136 interviewer utterances and

3702 respondent utterances). Table 4 shows the distribution of utterances across questions.

Table 4: Distribution of Utterances across Questions, Respondents and Interviewers

Speaker	**Question 1 (Nonsensitive, Complex)**	**Question 2 (Sensitive, Complex)**	**Question 3 (Sensitive, Noncomplex)**	**Question 4 (Nonsensitive, Noncomplex)**	**Question 5 (Income)**	**Total**
Interviewer	1210	563	815	1319	1229	5136
Respondent	857	401	547	931	966	3702
Total	2067	964	1362	2250	2195	8838

The distribution of utterances across questions, interviewers and respondents gives a picture of one facet of the interviewer-respondent interactions. Interviewers were responsible for more utterances than respondents (1434 more utterances overall). Questions 4 and 5 required the most utterances to complete, with Question 1 closely behind. Question 2 produced the least number of utterances, and Question 3 was between Question 2 and Question 1.

The nesting of utterances within questions is not modeled in this dissertation. Number of utterances per question is one measure of length of the interaction and can be used at the question level as a measure of affective burden (e.g., discomfort). Utterance-level coding (in addition to providing data for later analyses) was motivated by concerns about measurement (coding) accuracy. Forcing the application of a code to each meaningful utterance helps guarantee that small behaviors will be captured and summation of applied codes (calculated after coding is complete) will be more accurate relative to the application of codes at the question level. Coding at the utterance level also retains the temporal order of behaviors within each question. Coding directly at the question level is possible (e.g., how many requests for clarification were there in this question), but this approach threatens coding accuracy (e.g., coders missing actions or adding incorrectly) and loses the temporal sequence of events (e.g., we wouldn't know where the requests for clarification came in the sequence).

SCA records initially indicated to which category each respondent belonged. In the process of coding, several respondents whose income nonresponse type was miscoded were identified. In total 10 cases appeared to be mischaracterized by the SCA, at least for the purposes of the dissertation. These miscoded cases included 6 cases that were miscoded as bracketed respondents, but were actually dollar amount respondents and 4 cases that were coded as dollar amount respondents, but were actually bracketed respondents. Table 5 shows the breakdown of respondents by income response type after recoding. The reader can refer to Chapter 1 for a more complete discussion of these income nonresponse types. Chapter 4 discusses differences between respondents who answer income in each of these categories.

Table 5: Distribution of Income Nonrespondent Type

Income Respondent Type	**Number of Respondents**	**Percent of Respondents**
Income Nonrespondent	60	32.4
Bracketed Respondent	56	30.3
Dollar Amount Respondent	69	37.3
Total	185	100.00

Appendix G reports the univariate distributions of the coded and extracted variables. One immediate point of note is that most of the variables are severely positively skewed or have Poisson-like or zero-inflated distributions. This is due in part to the rarity of some of the utterance-level codes (e.g., requests for clarification, digression, negative comments). Other variables, while present on each utterance, are distributed in such a way that values on one extreme are common and values on the other extreme are rare (e.g., affect valence is usually neutral, affect intensity is usually low, and cognitive difficulty is usually not present). In Chapter 4, severely skewed variables will be recoded into binary indicators (e.g., 0 = none present, 1 = at least one present). This will allow for modeling the presence of indicators at each question, which avoids the challenge of dealing with the non-normal distributions of these continuous forms of these indicators.

Chapter 3

Sensitive and Complex Survey Questions and their Influence on Indicators of Affect and Cognitive Difficulty

3.1 Introduction

Characteristics of survey questions put demands on respondents. Some questions ask about personal topics that are emotionally demanding. Others require intense mental effort (e.g., memory or calculation) and are thus cognitively demanding. Some questions are demanding in both ways, and some in neither. When response is verbal, as it is in telephone surveys, we can study the way questions are answered to help understand the psychological processes that respondents experience as they attempt to arrive at an answer. The goal of the chapter is to explore the relationship between question characteristics (e.g., sensitivity and complexity) and respondents' verbal paradata that are indicative of respondents' psychological processes. Hypotheses are that sensitivity will increase indicators of affect and cognitive complexity will increase indicators of cognitive difficulty.

The data set produced in this study includes respondent behavior, voice, and speech (e.g., verbal paradata) at each respondent utterance within each of five questions. The first four of these questions are characterized as, 1) sensitive and complex, 2) sensitive and noncomplex, 3) nonsensitive but complex, and 4) nonsensitive and noncomplex. These characteristics were established by ratings of all questions in the SCA instrument. The fifth question asks about annual household income. Income questions are thought to be both sensitive and complex (See Chapter2 for text of the individual questions and discussion of the expert ratings).

Each of these questions is a repeated measure (e.g., each respondent has five repeated measures on each behavior, speech, and voice variable), where the measures are question-level summaries of utterance-level paradata as described in Chapter 2. Question characteristics (sensitivity and cognitive complexity) will be analyzed as two within-subjects factors with two levels each (presence or absence). Questions 2 and 3 are sensitive (1 and 4 nonsensitive), while questions 1 and 2 are complex (3 and 4 noncomplex). The sensitivity and complexity within-subjects factors define the analyses in this chapter.

3.2 Description of Question-level Data and Repeated Measures Analysis

Although data were coded on five questions including the income question, only the four questions before the income question are analyzed in this chapter. The goal of the chapter is to see how question sensitivity and complexity relate to respondents' question-answering behaviors, speech, and voice, restricting the analysis to questions where sensitivity and complexity are more clear-cut than they are for income questions (i.e., the four questions before income that were selected for these properties). Chapter 4 covers the relationship of behaviors, speech, and voice on questions prior to income with how respondents answer income. The analysis in the current chapter will show how question sensitivity and complexity influence indicators that are hypothesized to be measures of affect and cognition. It is generally expected that indicators of affect will be more prevalent on questions that are sensitive, while indicators of cognitive difficulty will be more prevalent on questions that are cognitively complex (see Chapter1 for specific hypotheses). This indicator-level exploration is a useful first step in exploring affective and cognitive dimensions of respondent verbal paradata using question characteristics as stimuli that should produce these indicators.

Two measures of each indicator can be constructed to summarize respondent behavior, speech and voice at each question. One involves taking a simple count of the event at each question. This can be defined as $\sum_{j=1}^{ik} B_{ijk}$, where the behavioral indicator (B) is coded at each of the j utterances within question k for respondent i, and j varies across individuals on each question simply because some respondents will take more utterances to answer the question. For the purposes of this study an utterance was defined in conjunction with the coding scheme to be the smallest unit of respondent speech to which a unique code (and only one code) could be applied. For example, if in their first conversational turn following the reading of the question, the respondent first says they don't know, and then asks for the question to be re-read, this conversational turn would be broken into two utterances. The first utterance would be given a code for the don't know response, and the second utterance would be given a code for the re-read request. Additional indicators that do not define utterance boundaries (e.g., ratings of affect and cognitive difficulty) are applied to each of these two utterances. Within-question variability will not be modeled in the dissertation.

Using counts, the summary of utterance-level codes that are analyzed as the dependent variable in the ANOVA is defined by a total at each question, or

$$\bar{y}_k = \frac{\sum_{i=1}^{n} y_k}{n},$$

where the numerator is the count of the indicator on each question for each respondent, or $y_k = \sum_{j=1}^{ik} B_{ijk}$, and the denominator is the sample size, which is the same at each question (e.g., $n = n_k$). The mean $\bar{y}_k$ is thus an average across respondents (*n*) of the sum of the indicator (*B*) over utterances (*j*) at each question (*k*).

A second method of constructing the question-level indicator takes the mean of each indicator over the number of respondent utterances (rather than the sum), so that the resulting question-level measure is the average indicator value per utterance. In the case of binary indicators, this is the proportion of utterances within the question having that indicator. The form of this indicator, designated here as a mean for question *k* is

$$\bar{y}_{ik} = \frac{\sum_{j=1}^{ik} B_{ijk}}{n_{ik}},$$

where *j* is the respondent utterance within a question, *i* is the respondent, and *k* is the question. Thus, B_{ijk} is the behavioral indicator for respondent *i* on utterance *j* of question *k*, and n_{ik} is the number of utterances for respondent *i* at question *k*. The means compared in the repeated measures ANOVA are then averages across respondents within each repeated measure (i.e., *k* questions)

$$\bar{y}_k' = \frac{\sum_{i=1}^{i} \bar{y}_k}{n},$$

where *n* is the number of respondents and $\bar{y}_k{}'$ is the mean of the average number of (or proportion of) times an individual indicator occurs at the question across respondents at question *k*.

Each of these two definitions of the question-level indicators presents a unique challenge to analysis and interpretation. When the indicator is defined as the total count of behavior at each question there is potential for the count of the behavior to be positively correlated with the length of the interaction on the question simply due to arithmetic. Indicators for which a value is assigned at each utterance by definition (e.g., affect intensity, affect valence, and cognitive difficulty ratings) will certainly be affected by the number of utterances if a sum is used. For binary indicators, longer questions give more opportunity for the indicator to occur, which may artificially inflate the total number of occurrences of the behavior, but reduce the proportion as the number of utterances increases. The inflation of counts on some variables is empirically verified by the correlations presented in Appendix H, though the range of correlation is wide (r=.035 to .720). The correlations in Appendix H support the plan to use means/proportions of indicators at each question, rather than counts of indicators.

Defining the question-level indicator as an average occurrence of the indicator at the question (or the proportion of respondent utterances on which the behavior occurs) removes any artificial correlation with interaction length but introduces other problems. When these averages and proportions are calculated, the meaning of the scale (i.e., proportion of utterances within the question exhibiting indicator *B*) can be difficult to interpret. A low proportion (i.e., .10), would most likely indicate few instances of the indicator and a high number of utterances (i.e., small numerator, large denominator).

Such a proportion has a relatively straightforward interpretation; the indicator was infrequent on that question. However, complications of interpretation come at the middle and upper end of the scale. A proportion of .5 could be obtained from 1 occurrence out of 2 utterances, or from 8 occurrences out of 16 utterances. The first is a quick exchange with only one instance of a problem. The other is a longer exchange (which may itself indicate a problem) with several instances of the problem indicator. For some indicators, 8 occurrences would suggest something quite different than 1 occurrence. For example, a respondent who requests clarification or a repeat 8 times would be expressing much more cognitive difficulty than one who asks for clarification or a repeat once (with 2 total utterances). These two cases would be considered equal under the mean/proportion definition. Furthermore, the lowest number of respondent utterances possible in a question is 1 (answer). For variables like “answers with qualifications”, “refusals”, and “don’t know” responses, it’s possible to have a proportion of 1.0. Although an answer with qualification may seem to indicate difficulty, if the interviewer accepts the answer, there might be only one respondent utterance (e.g.,. proportion of 1.0). Thus a high level of “answering with qualification” could indicate a lot of trouble (if the exchange is also long) or little or no trouble (if the exchange is short).

The decision of which indicator to use is partly a conceptual one. Aside from the correlation of some indicators with number of utterances (a statistical issue), either indicator is statistically acceptable, but the appropriateness depends on whether the research question is about modeling the “the total occurrence (count) of each behavior” or “the proportion of utterances (average) in the question that exhibit the behavior.” One is a measure of presence and frequency, the other a measure of percentage of difficult (or ease). While 8 counts of request for clarification (of 16 utterances) may be qualitatively different from 1 of 2, it is still reasonable to say that each of these questions was “50% difficult” for the respondent answering. Also, we do not expect to see such extreme cases as 8 requests for clarification. Given this conceptual justification, low-risk for measurement error, and the absence of specific hypotheses about counts of behavior, averages and proportions are the best solution for this chapter.[10]

Another problem with these indicators, is that many of these indicators are rare events with most respondents having no occurrence at the question and few having a count greater than one (i.e., a proportion or utterances greater than 0). This often leads to seriously skewed distributions that cannot be corrected by transformations. All data are treated as continuous without transformations in this chapter.[11]

The multivariate model is used for analysis of the within-subjects factors (sensitivity and complexity) in the repeated measures ANOVAs presented in this chapter. Using the multivariate approach allows relaxation of the requirements of compound symmetry and sphericity implied by the univariate model (Keppell & Wickens, 2004). The within-subjects results presented in this chapter are all based on multivariate model

[10] To explore any substantive differences in findings created by these different definitions of the question-level indicator, counts and proportions were both analyzed with repeated measures ANOVA. Results are almost identical with respect to presence and direction of differences, so only proportions and means are presented in the dissertation.

[11] For severely skewed distributions with high proportions (e.g., 80% of more) of zeros, power transformations (inverse, square root, etc) did not improve distributions.

F-values and p-values. In most instances, the substantive interpretation of multivariate and univariate ANVOA results are identical.[12]

3.3 Effects of Question Sensitivity and Complexity on Respondent Verbal Paradata

Appendix I presents a table of F-values and p-values of main effects and interactions for the multivariate ANOVA model with each indicator as the dependent variable. All indicators analyzed are included in this table, while only significant effects are discussed in the text.

3.3.1 The Effect of Question Sensitivity

One goal of this analysis was to determine the effect of question sensitivity on indicators of affect and cognitive difficulty. It was expected that indicators hypothesized to be measures of affect would be influenced by question sensitivity, but sensitivity may also impact measures of cognitive complexity in ways that were not hypothesized. Specifically, the expectation was to see higher rates of indicators of affect (e.g., refusals, affect ratings, voice pitch) on items that are higher in sensitivity and no effect of question sensitivity on indicators of cognitive difficulty (e.g., reports, repairs, cognitive difficulty ratings, fillers). The impact of sensitivity is reviewed, first on indicators of affect and then on indicators of cognitive difficulty. Table 6 presents those indicators that have significant main effects of sensitivity at the $\alpha=.05$ significance level. The models tested included main effects for question sensitivity and question complexity, and interactions

[12] Reviews of the covariance structure of the repeated measures confirms that the multivariate model is the more appropriate model for many of the indicators, even if substantive results do not differ from the univariate repeated measures model.

of these two factors. The fifth column in Table 6 reports whether a significant interaction (α=.05) was present for that indicator. Main effects should be interpreted with caution for indicators on which there is an interaction with cognitive difficulty. Interactions are addressed later in the chapter.

Table 6: Significant Effects of Sensitivity on Indicators (Repeated Measures ANOVA w/ Four Items before Income, α=.05 level only)

Construct	Variable[13]	Direction of Difference	Difference	Interactions and Other Effects
Affect	Question length in utterances (count)	Longer in nonsensitive questions	5.408 (<.0005)	Complexity effect
	Overspeech	More in nonsensitive questions	.061 (.007)	Complexity effect
	Backchannel	More in nonsensitive questions	.013 (.007)	No other effect
	Conversation management	More in nonsensitive questions	.006 (.045)	No other effect
	Agreement	More in nonsensitive questions	.016 (.008)	No other effect
	Affect Intensity (0-9 at each utterance)	Higher in nonsensitive questions	3.376 (<.0005)	Complexity effect; Interaction w/ Complexity
Cognitive Difficulty	Duration of first respondent utterance	Longer in nonsensitive questions	1.675 (<.0005)	Complexity effect; Interaction w/ Complexity
	Fillers per utterance	More in nonsensitive questions	.155 (.004)	Complexity effect; Interaction w/ Complexity
	Filler duration per utterance	Longer in nonsensitive questions	.625 (.036)	Complexity effect; Interaction w/ Complexity
	Pauses per utterance	More in nonsensitive questions	.279 (<.0005)	Complexity effect; Interaction w/ Complexity
	Pause duration per utterance	Longer in nonsensitive questions	4.233 (<.0005)	Complexity effect; Interaction w/ Complexity
	Respondent words per utterance	More in nonsensitive questions	4.705 (<.0005)	Complexity effect; Interaction w/ Complexity
	Answers primary question	More in sensitive questions	.276 (<.0005)	Complexity effect; Interaction w/ Complexity
	Answers primary question with qualification	More in sensitive questions for	.086 (<.0005)	Complexity effect; Interaction w/ Complexity
	Uncertain about answer	More in nonsensitive questions	.015 (.001)	No other effects
	Explicit don't know	More in nonsensitive questions	.025 (.001)	Interaction w/ Complexity
	Implied don't know	More in nonsensitive questions	.006 (.036)	No other effects
	Repair only	More in nonsensitive questions	.034 (.001)	No other effects
	Stammer only	More in nonsensitive questions	.033 (.004)	Interaction w/ Complexity
	Repair and stammer	More in nonsensitive questions	.037 (<.0005)	Complexity effect; Interaction w/ Complexity

[13] Unless otherwise noted, each indicator is defined as the proportion of utterances on the question that have that indicator, or the average over respondent utterances within the question.

Significant sensitivity effects were found on the following indicators of affect: length of questions in utterances, overspeech, backchannels, conversation management, agreement, and affect intensity. The significant effects of question sensitivity on measures of affect show a trend of higher rates of affect indicators on nonsensitive items, counter to hypotheses. Affect intensity followed the same pattern. Affect was less intense on sensitive questions, counter to hypotheses.

Alternative mechanisms warrant exploration given these unexpected findings. Question sensitivity clearly places psychological demands on respondents, but rather than producing more indicators of affect, it produces fewer. The most parsimonious explanation for this finding is that a question's sensitive content uses psychological resources (e.g., determining if the question is too sensitive to answer, or whether to misreport) that would otherwise be used communicating with the interviewer. When those resources are occupied with demands of the question, fewer are available for communication. Most of the affect indicators (e.g., number of utterances, overspeech, backchannels, conversation management, and agreement) can also be interpreted as indicators of conversationality, and respondents' attention to producing conversational cues may be subdued when contemplating a particularly sensitive question. A more complex and more socially-oriented interpretation assumes that the reduced communication is conscious. When respondents are threatened by question content, respondents may intentionally reduce any behavior that would encourage the interviewer to talk more about the threatening content (e.g., backchannels, fillers). In addition to being less conversational, respondents seem to be more emotionally withdrawn on sensitive items. Coders rated them as having less affect intensity on questions that were sensitive, compared to nonsensitive questions. It was expected that sensitivity would heighten affect intensity, for example, perceived shock or nervousness as a response to the threatening content, but that was not found. It may be that respondents emotionally withdraw from the interaction either due to limited cognitive resources or intentionally to avoid having to talk more about the topic. It is also likely that conversationality and emotional involvement (affect intensity) are part of the same dimension. When people talk more, they also tend to be more affectively involved (i.e., interested in the conversation) and vice versa. This seems to be supported by the data. These alternative mechanisms will recur throughout interpretation of the findings.

Effects of sensitivity were also found on indicators of cognitive difficulty, including duration of the first respondent utterance, fillers per utterance, filler duration per utterance, pauses per utterance, pause duration per utterance, respondent words per utterance, answering the question (with and without qualification), uncertainty about the answer, explicit and implied “don’t know” responses, repairs, and stammers. Like the effect on affect indicators, question sensitivity reduced most indicators of cognitive difficulty (i.e., more were found on nonsensitive questions), supporting alternative mechanisms such as a psychological resource allocation. Answering the question (with and without qualification) was the only cognitive difficulty indicators that were more present on sensitive questions. Higher rates of question-answering are evidence that higher proportions of utterances are spent answering the question. This could be due to more utterances on which answers are given (with our without qualification), which would be a sign of trouble because the respondent is trying to answer but the interviewer is not accepting it. It could also be due to more paradigmatic exchanges (e.g., less trouble; Maynard & Schaeffer, 2002b; Schaeffer & Maynard, 2002). A question on which the only respondent utterance is an answer would be coded as having 100% of utterances on which the question was answered. Fewer utterances were found on sensitive questions, and thus short and more paradigmatic exchanges seem plausible. Paradigmatic (i.e., quicker) responses on sensitive questions (i.e., higher proportion of utterances in which the question is answered) would fit with an explanation citing intentionally-reduced conversationality, rather than a cognitive resources explanation, in which respondents are trying to complete sensitive questions as quickly as possible.

No effect of question sensitivity was hypothesized for indicators of cognitive difficulty, but some effects found. In retrospect, these indicators may mean something different on sensitive questions than they do on cognitively difficult questions. If respondents are more comfortable or less distracted by question demands on nonsensitive questions, and are thus more conversational, these “indicators of cognitive difficulty” would likely increase as well. In other words, verbal paradata that indicate cognitive difficulty when they are present on complex questions may indicate cognitive resource expenditure or discomfort when absent in threatening questions.

3.4 Effects of Question Complexity

Table 7 presents indicators on which main effects of complexity where found, and the direction and amount of the difference. As with sensitivity effects, hypothesized, non-hypothesized and counter-to-hypothesis effects were found on affect indicators and cognitive difficulty indicators. The same caution as noted in Table 6 should be taken with interpreting main effects when interactions are present.

Table 7: Significant Effects of Cognitive Difficulty on Indicators (Repeated Measures ANOVA w/ Four Items before Income, α=.05 level only)

Construct	**Indicator**	**Direction of Difference**	**Difference in Means of Proportions**	**Interactions and Other Effects**
Affect	Laughter	More in noncomplex questions	.018 (.031)	Interaction w/ Sensitivity
	Average affect intensity	Higher in noncomplex questions	.900 (.001)	Sensitivity Effect
	Average affect valence	Higher in noncomplex questions	.154 (.001)	No other effects
Cognitive Difficulty	Question length in utterances	Longer in noncomplex questions	1.559 (<.0005)	Sensitivity Effect; Interaction w/ Sensitivity
	Overspeech per utterance	More in complex questions	.081 (.001)	Sensitivity Effect
	Fillers per utterance	More in noncomplex questions	.279 (<.0005)	Sensitivity Effect; Interaction w/ Sensitivity
	Filler duration per utterance	Longer in noncomplex questions	1.797 (<.0005)	Sensitivity Effect; Interaction w/ Sensitivity
	Pauses per utterance	More in noncomplex questions	.310 (<.0005)	Sensitivity Effect; Interaction w/ Sensitivity
	Pause duration per utterance	Longer in noncomplex questions	5.284 (<.0005)	Sensitivity Effect; Interaction w/ Sensitivity
	Total words per utterance	More in complex questions	2.021 (<.0005)	Interaction with Sensitivity
	Respondent words per utterance	More in noncomplex questions	3.470 (<.0005)	Sensitivity Effect; Interaction w/ Sensitivity
	Duration of first respondent utterance	Longer in noncomplex questions	1.718 (<.0005)	Sensitivity Effect; Interaction w/ Sensitivity
	Answers primary question without qualification	More in complex questions	.093 (<.0005)	Sensitivity Effect; Interaction w/ Sensitivity
	Answers primary question with qualification	More in complex questions	.056 (.002)	Sensitivity Effect; Interaction w/ Sensitivity
	Requests clarification or repeat	More on complex questions	.015 (.039)	No other effects
	Repair and stammer	More in noncomplex questions	.020 (.014)	Sensitivity Effect; Interaction w/ Sensitivity

Laughter was the only objective indicator of affect that was affected by question complexity. Although no effect of complexity on affect indicators was hypothesized, an increase of laughter due to complexity might suggest that laughing stems from discomfort with the complexity of the question (e.g., nervous laughter). Yet the actual result (more laughter on noncomplex questions), is counter-intuitive based on theory and mirrors the effects of question sensitivity previously discussed. Affect intensity and valence follow the same pattern. Intensity was higher on noncomplex questions, suggesting more perceived feeling on these questions. Affect valence was also higher and positive on noncomplex questions. These two results, combined with the finding for laughter, are evidence that respondents feel better while answering noncomplex questions than complex ones.

Cognitive resource and conversationality mechanisms may also relate question complexity to indicators of affect, similar to the relationship proposed between question sensitivity and verbal paradata. That is, when questions are complex, respondents may be so focused on comprehending the question and calculating an answer that all conversational and affective indicators are reduced. When questions are less cognitively demanding, respondents may be freer to engage in conversational behavior like laughter. These findings do not support the idea that respondents laugh as way to lighten cognitive demands. Rather, they laugh less and affectively engage less when cognitive demands are high.

Unlike the effect of complexity on affect indicators, the effects of question complexity on indicators of cognitive difficulty are mixed. Some show higher prevalence on noncomplex items following the trend observed thus far. These include the number of utterances, filler presence and duration, pause presence and duration, respondent words per utterance, duration of the first respondent utterance, and repairs and stammers. Each of these indicators reflects more talking, whether intentional or not, and their increased presence on noncomplex questions is likely explained by cognitive resource and conversationality mechanisms discussed already. Even pauses, which are empty space, are signs of more conversation. The pauses analyzed here were all within-utterance pauses, during which respondents are holding the floor. They may be actively thinking of an answer, or simply drawing out the length of the conversation. Fillers show the same pattern as pauses, and may also be used consciously by respondents to hold the floor while talking. If fillers are used to retain a conversational turn, fewer would be expected on complex questions that the respondent wants to finish as quickly as possible (e.g., complex questions). This is what was found.

Other indicators of cognitive difficulty show a hypothesized relationship with question complexity. Overspeech, total words per utterance, answering (with and without qualification), and requests for clarification or repeat were higher on cognitively complex questions. As expected, complexity increases requests for clarification or a repeat of the question. Complex questions are harder for respondents to understand, and lead to these requests. Overspeech is more frequent as well, perhaps coinciding with requests for clarification. This conversational obstacle is expected if respondents are having a hard time understanding the question or deciding how to answer.

Answering with or without qualification is also higher on complex questions than noncomplex ones. This was seen as a response to question sensitivity as well, with sensitive questions producing a higher proportion of utterances on which the respondent answers, possibly due to more paradigmatic and quicker exchanges. This explanation can be applied here. When questions are difficult respondents attempt to answer and move on as quickly as possible.

Finally, total words per utterance is higher on complex questions than noncomplex questions. This is only interesting because respondent words per utterance are lower on complex questions, suggesting that the difference in effects may be due to interviewer words. While respondents were talking less on complex questions, interviewers may be talking more, perhaps due to the length of the question, or in an effort to help respondents come to an answer.

Summarizing across indicators and question characteristics, the common theme is that hypothesized indicators of sensitivity and complexity were higher on questions with fewer affective or cognitive demands. If we take these indicators as measures of conversationality, a purpose to which they all seem to apply well, we can infer that respondents are more conversational on items that are lower in demand. The question remains, “why are respondents more conversational when question demands are lower?” The most parsimonious explanation is that when question demands are high, whether in affective (sensitivity) or cognitive (complexity) content, respondents’ cognitive resources will be reduced, limiting their ability to engage in conversation. Respondents may be so absorbed by the question that extraneous conversational cues are reduced. A more social explanation would suggest that they are intentionally reducing conversational cues in an effort to move past the threatening or challenging question. The next section will look at interactions between sensitivity and complexity to see how these factors work together to produce respondent verbal paradata.

3.5 Interactions between Sensitivity and Complexity

Interactions between sensitivity and complexity were found on explicit refusals, laughter, speech rate, pitch range, pitch standard deviation, duration of first respondent utterance, answering with and without qualification, expressions of uncertainty about the question, explicit don't knows, digressions without an answer, reports, stammers, repairs and stammers, fillers (presence and duration), pauses (presence and duration), total words per utterance, and respondent words per utterance.

Explicit refusal, laughter, speech rate, and pitch were all indicators of affect that showed interactions of question sensitivity and complexity. Laughter, pitch variability and pitch range show a pattern that can be explained by the inverse relationship between question demands, cognitive resources, and conversationality that was introduced above. The highest rates of these indicators of affect were found on the nonsensitive, noncomplex question, further supporting the hypothesis that the absence of question demands increases verbal behavior. The interaction on each of these three indicators is such that there was no statistically significant difference (α=.05) between levels of question complexity when questions were sensitive, but a significantly higher prevalence was found on the noncomplex question when questions were also nonsensitive (i.e., all were ordinal interactions). Figures 2-4 show these interactions.

Figure 2: Interaction of Sensitivity and Complexity on Laughter

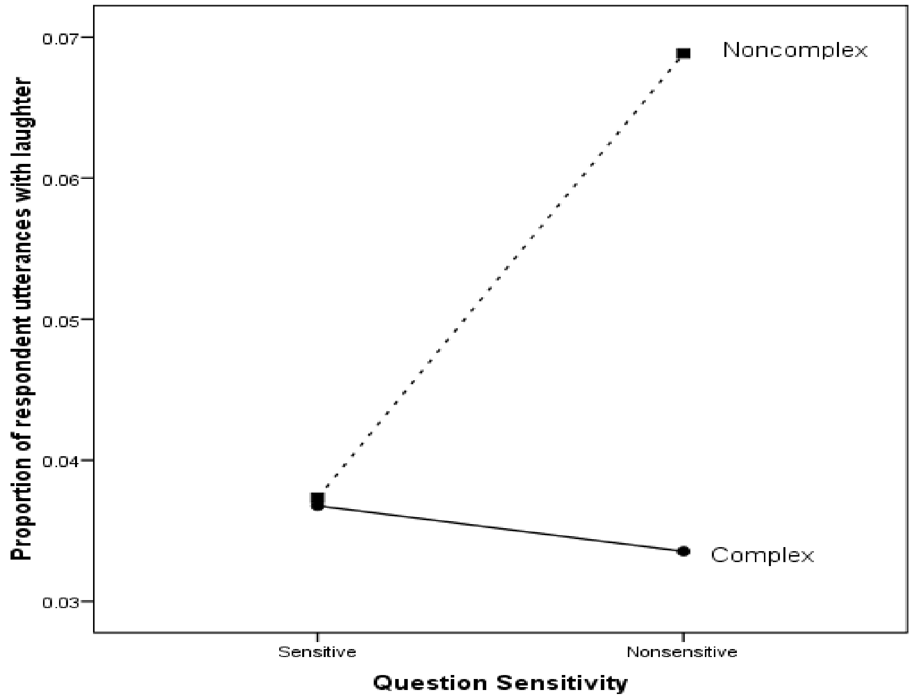


Figure 3: Interaction of Sensitivity and Complexity on Pitch Standard Deviation

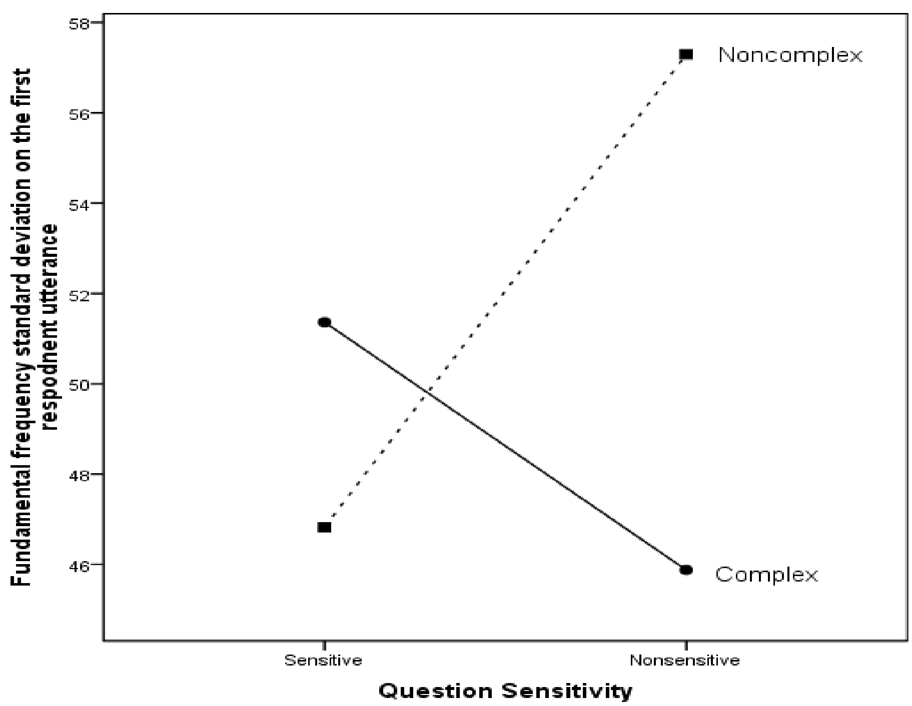
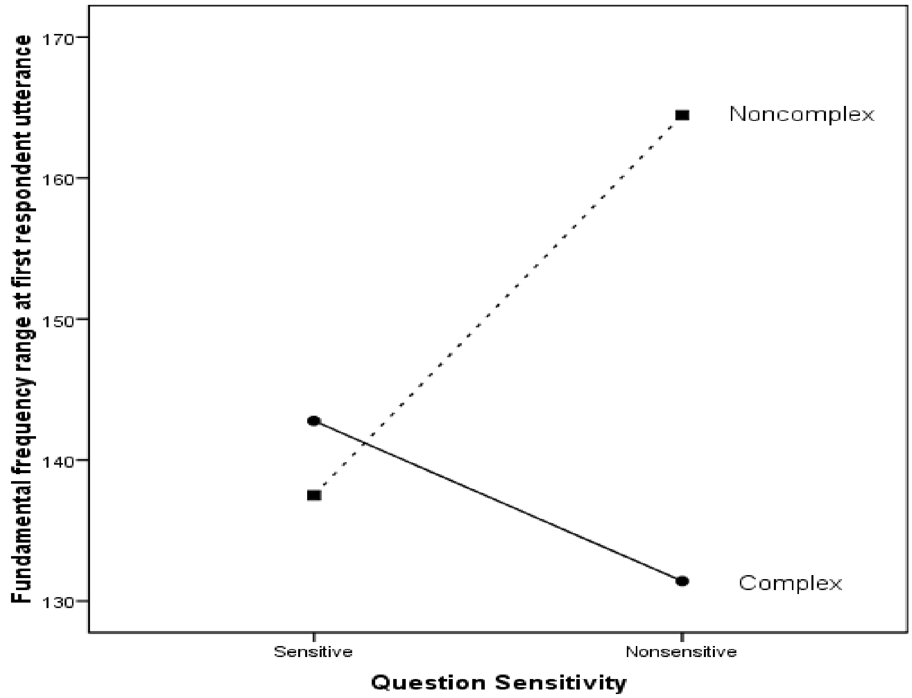
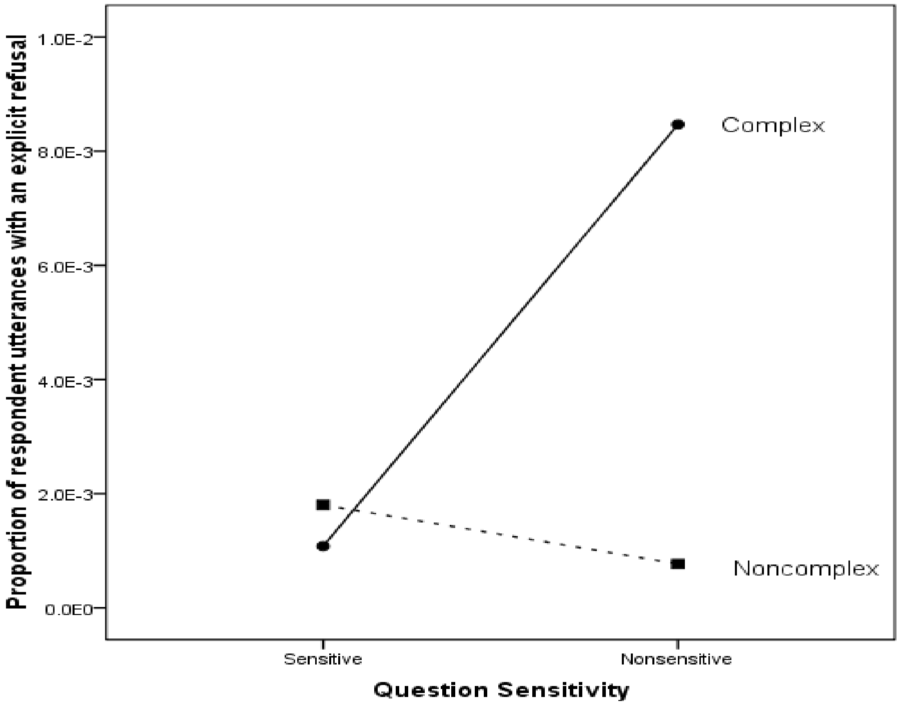


Figure 4: Interaction of Sensitivity and Complexity on Pitch Range



Refusals show a pattern in which the highest rate of explicit refusal was found on the nonsensitive, complex question. When questions were sensitive, there was no significant difference (α=.05) between levels of complexity, similar to what was found on other affect indicators. The effect of complexity only emerges when questions are also nonsensitive (i.e., the interaction is ordinal), and it is in the opposite direction of affect indicators reviewed so far (higher on the complex, nonsensitive question). This result is initially counter-intuitive, as more refusals would be expected on sensitive items, not complex ones. However, respondents may also refuse to answer questions that are too difficult to answer, and so a higher rate of refusals is not completely surprising. Figure 5 shows the interaction of sensitivity and complexity on laughter and refusals respectively.

Figure 5: Interaction of Sensitivity and Complexity on Explicit Refusals

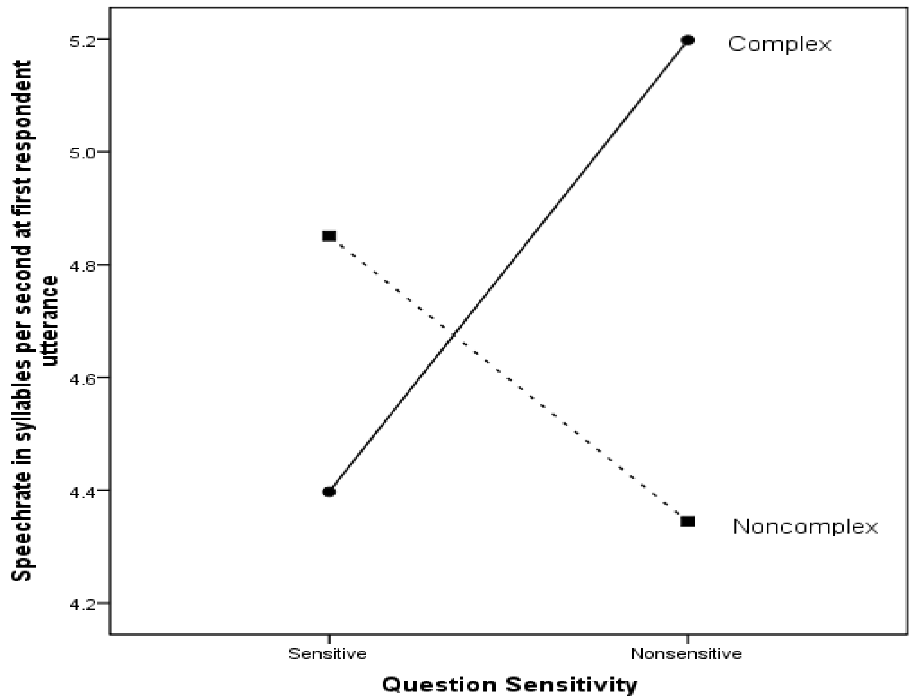


Speech rate showed a true disordinal[14] interaction, with differences between levels of question complexity when items were both sensitive and nonsensitive. When items were sensitive, the noncomplex item had the fastest speech. When items were nonsensitive, the complex item produces the fastest speech. It seems that either one question demand or the other (sensitivity or complexity) will increase speech rate, perhaps in an effort to complete the question as quickly as possible. Figure 6 shows this interaction.

---

[14] Disordinal interactions (cross-over interactions) are those in which there are differences between levels of cognitive difficulty at each level of sensitivity, and the direction is reversed at each level. Ordinal interactions are those for which there is no difference between levels of complexity for one level of sensitivity, but differences between levels of complexity at the other. Each interaction is indicated as being disordinal or ordinal so the reader can determine which mean differences presented in the figure are statistically significant.

Figure 6: Interaction of Sensitivity and Complexity on Speech Rate



Hypothesized indicators of cognitive difficulty also showed interactions of sensitivity and complexity (answers with and without qualification, uncertainty about the question, explicit don't knows, digressions, reports, stammers, repairs, filler presence and duration, pause presence and duration, total words, and respondent words). Most of these showed an interaction pattern that fits a reduced cognitive resource or reduced conversationality explanation. The highest rates were found on the nonsensitive, noncomplex question with no difference between levels of complexity when questions were sensitive. This pattern was found for uncertainty about the question, fillers per utterance, filler duration per utterance, pauses per utterance, pause duration per utterance, respondent words per utterance, stammers, repairs and stammers, and digressions with no codable answer. Figures 7-15 show these interactions.

Figure 7: Interaction of Sensitivity and Complexity on Uncertainty about the Question

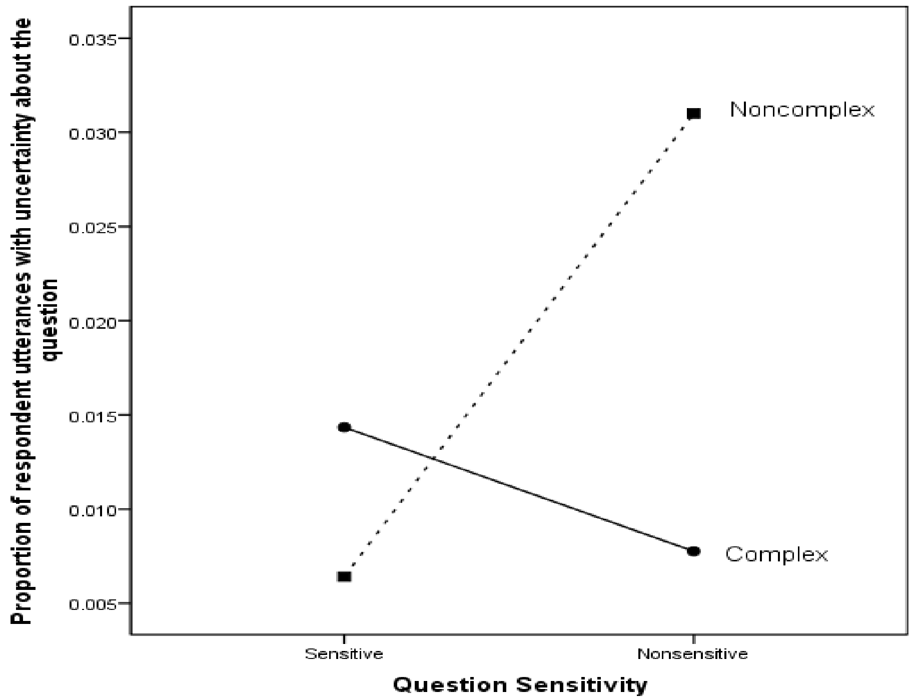


Figure 8: Interaction of Sensitivity and Complexity on Fillers per Utterance

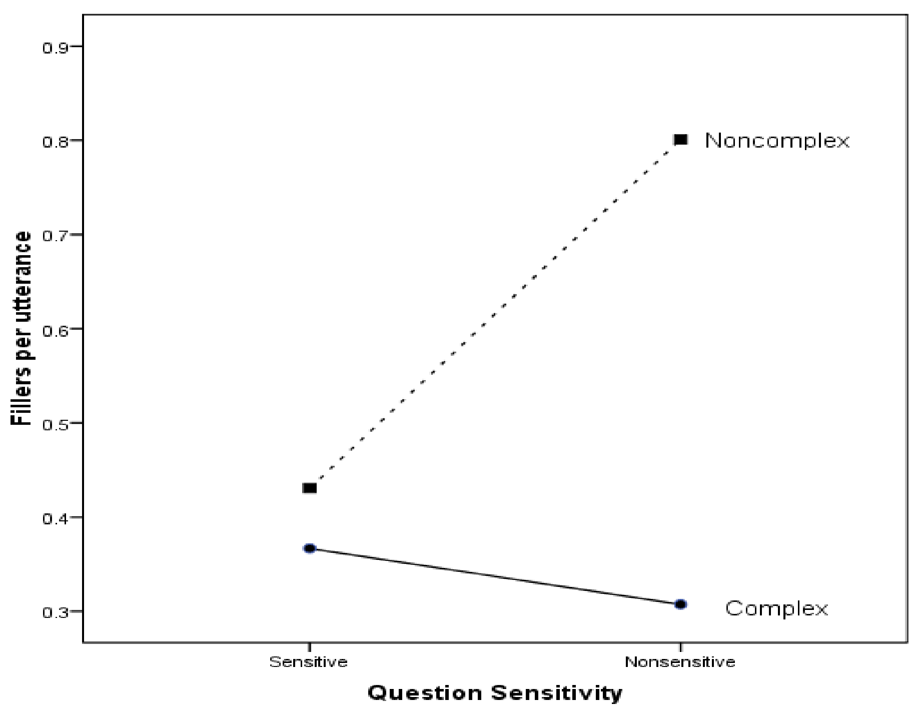


Figure 9: Interaction of Sensitivity and Complexity on Filler Duration per Utterance

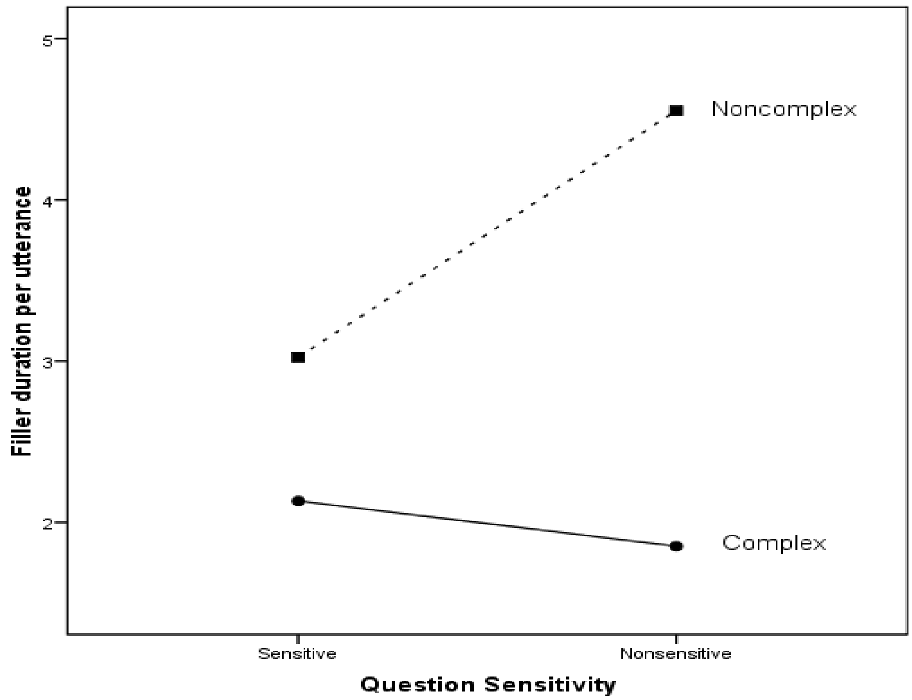


Figure 10: Interaction of Sensitivity and Complexity on Pauses per Utterance

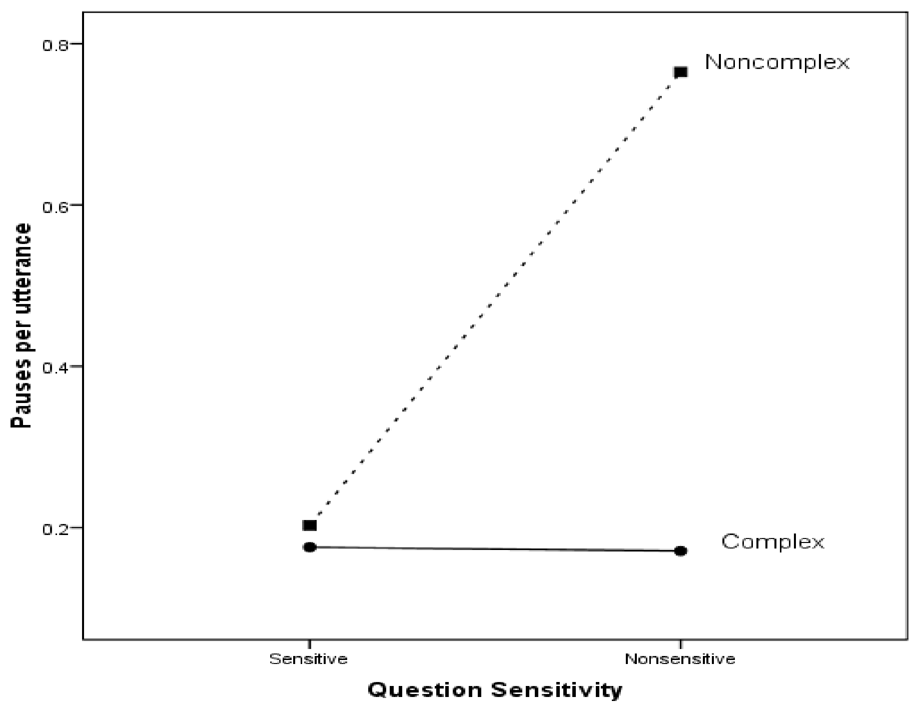


Figure 11: Interaction of Sensitivity and Complexity on Pause Duration per Utterance

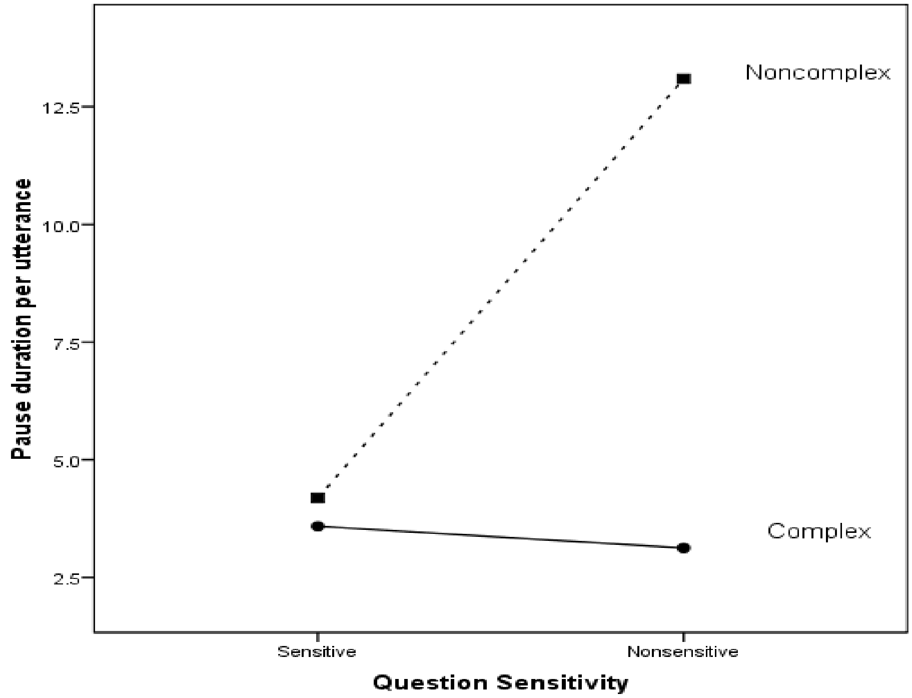


Figure 12: Interaction of Sensitivity and Complexity on Respondent Words per Utterance

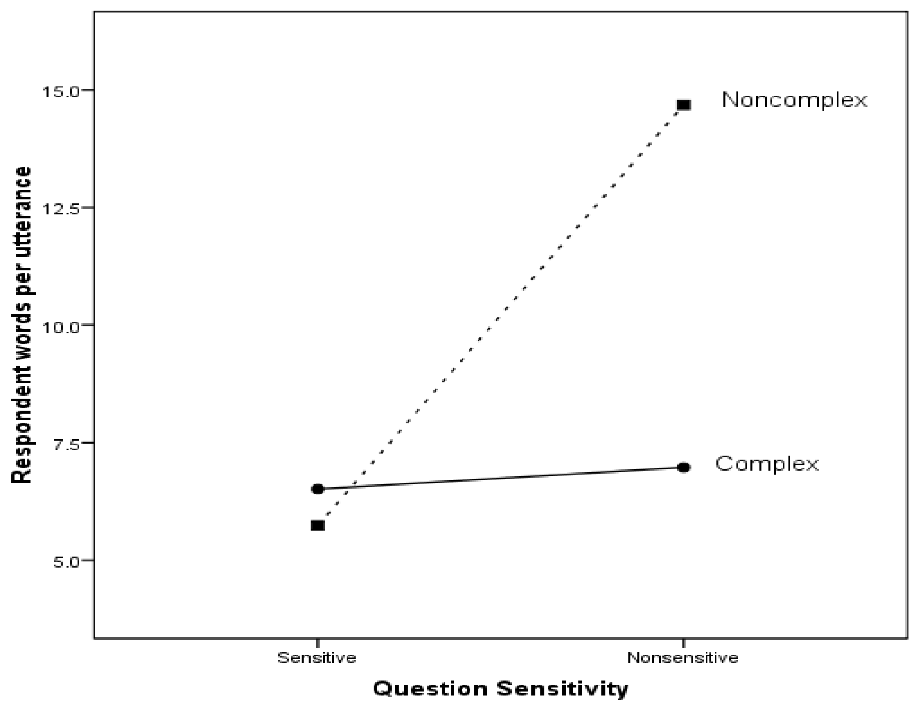


Figure 13: Interaction of Sensitivity and Complexity on Stammers Alone

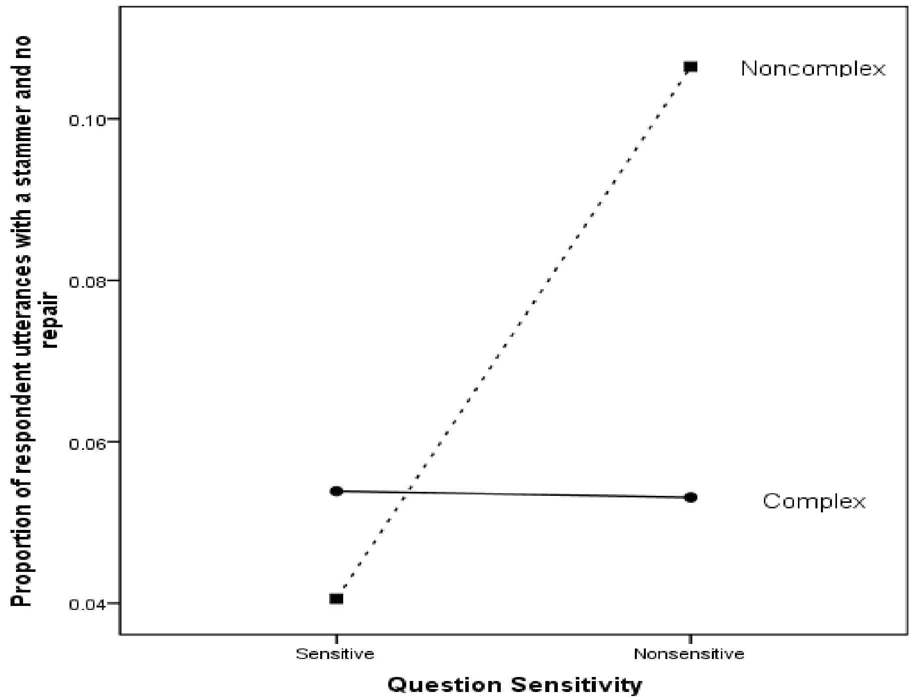


Figure 14: Interaction of Sensitivity and Complexity on Repairs and Stammers

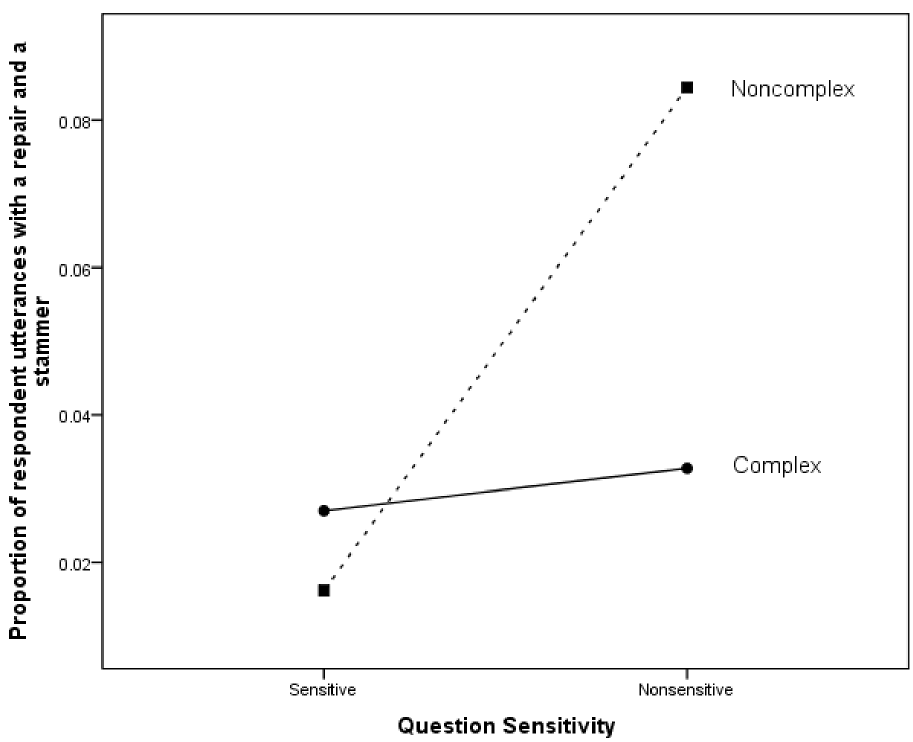
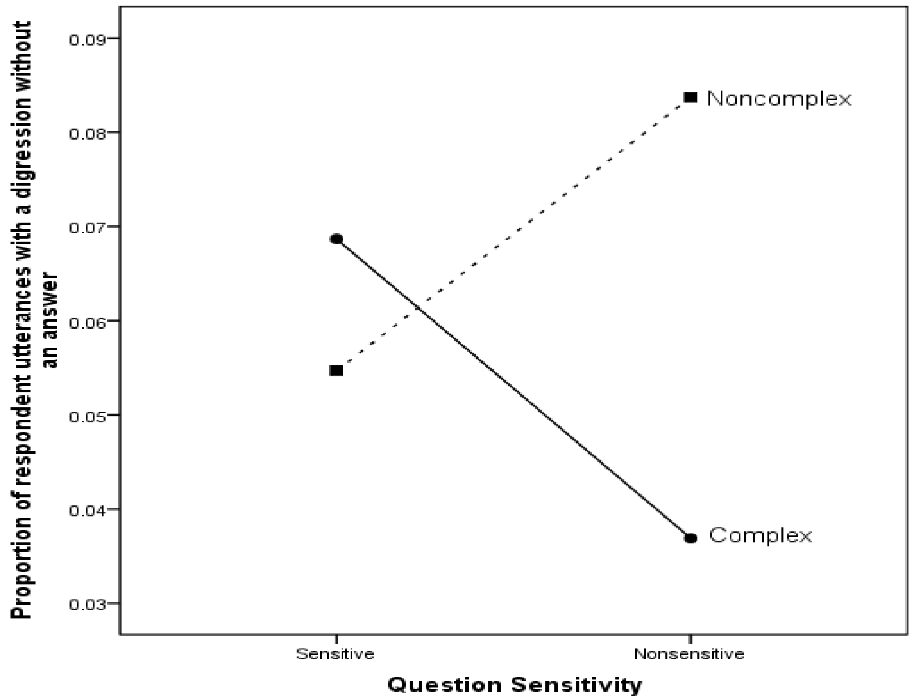
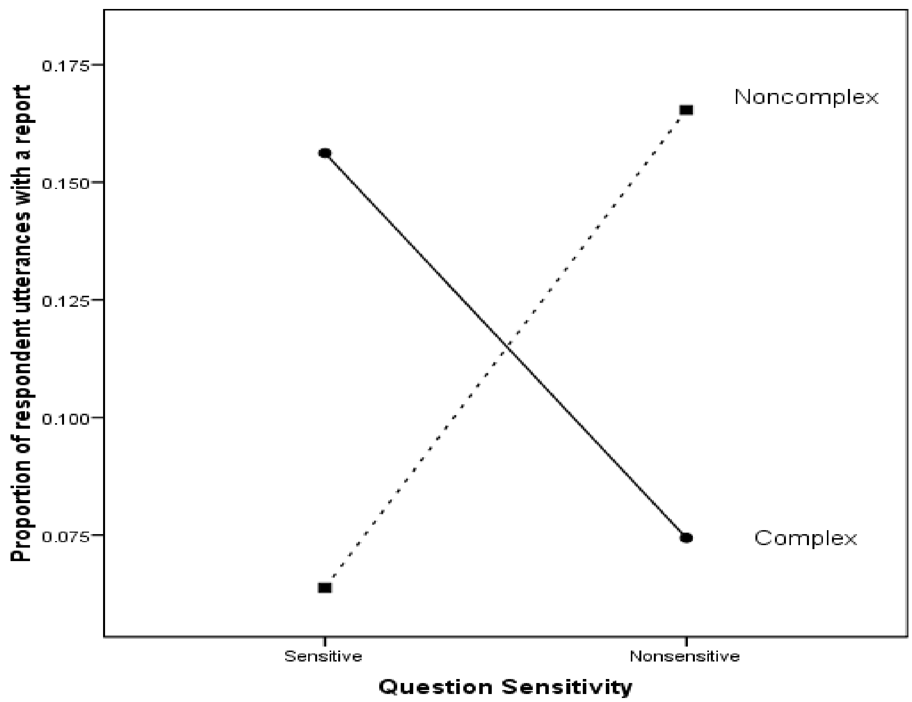


Figure 15: Interaction of Sensitivity and Complexity on Digressions with No Answer



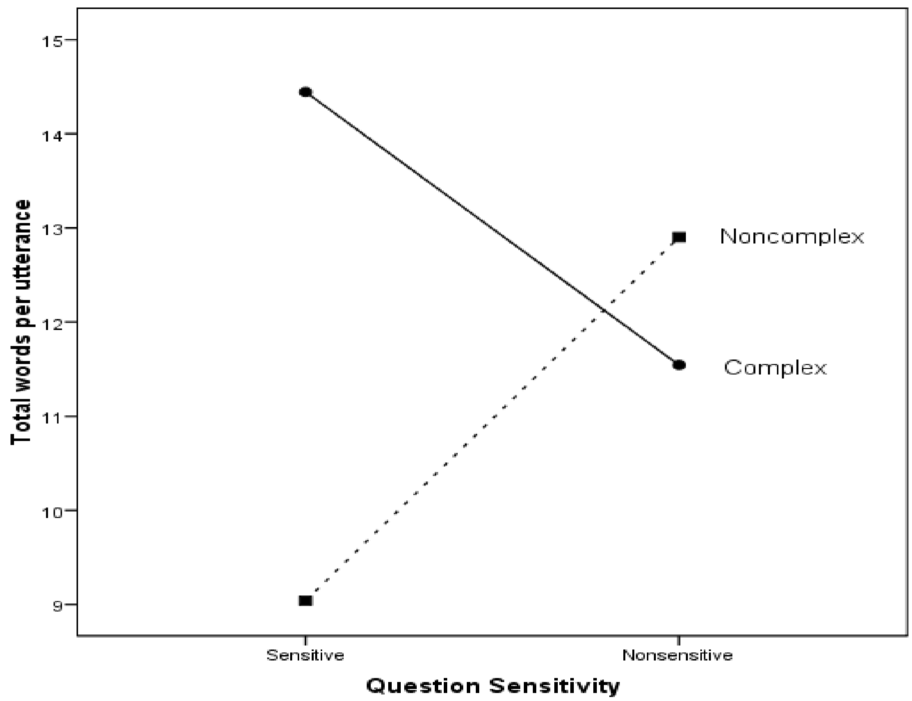
Reports show a disordinal interaction between sensitivity and complexity in which the nonsensitive, noncomplex question and the sensitive, complex question have the highest rates of reports (see Figure 16). Reports are generally thought to be caused by uncertainty about how to answer questions, a type of cognitive difficulty (Schober & Bloom, 2004), but it is possible that reports also reflect affective reactions to questions (e.g., providing only vague question-relevant information that does not explicitly reveal information deemed by the respondent to be too personal to share with the interviewer, and thus does not answer the question). While this alternative explanation of reports was not anticipated, it provides an explanation for how affect might influence reports. What is surprising is that they are also highest on the question that is neither complex nor sensitive. We could interpret this as increased conversationality in the absence of question demands if reports are taken as a sign of conversationality. The interaction seems to provide contradictory evidence; 1) reports as signs of trouble occur more when questions are both sensitive and complex, and 2) reports as signs of conversationality occur more when questions are neither sensitive nor complex. These interpretations seem incompatible, and it is not immediately clear what explanation would reconcile this result outside of the role of idiosyncratic question characteristics that are not evaluated here.

Figure 16: Interaction of Sensitivity and Complexity on Reports



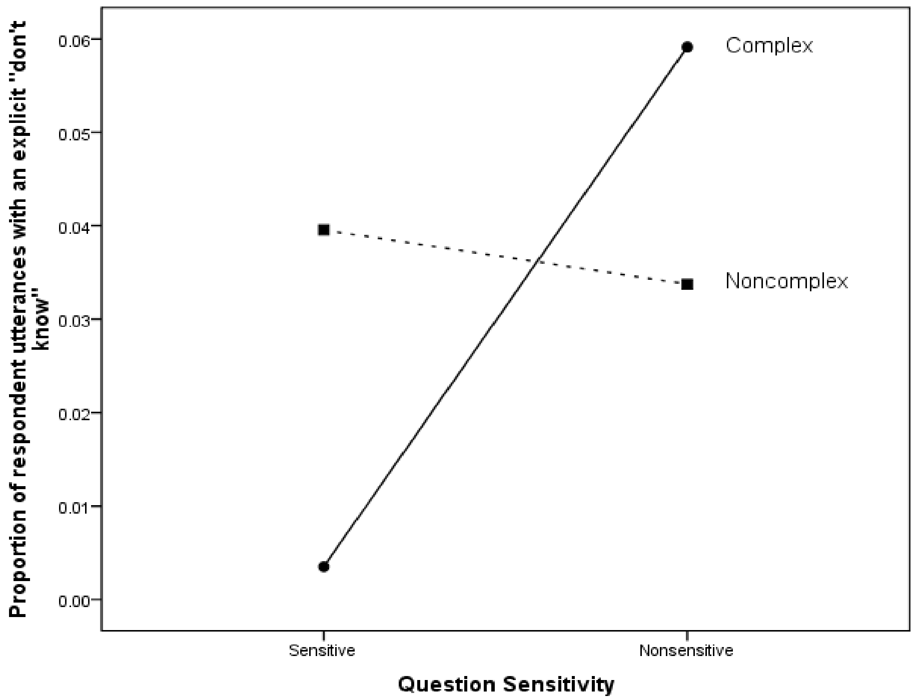
Total words per utterance show an interaction that is opposite of the general interaction trend. For total words, differences between levels of complexity were found only when questions were also sensitive. The highest rates are found on sensitive, complex questions. Compare this to respondent words per utterance (and most other indicators) in which differences between levels of complexity are only present when questions are nonsensitive, and the highest rate of total words per utterance was found on the nonsensitive, noncomplex question. Total words include interviewer words as well as respondent words, and would include reading of the question.

Figure 17: Interaction of Sensitivity and Complexity on Total Words per Utterance



Explicit don't know responses showed a disordinal interaction of sensitivity and complexity. The same rate of don't knows was found when questions were noncomplex, regardless of sensitivity. For complex questions, however, there was a much higher rate of explicit don't knows when questions were nonsensitive and complex (compared to sensitive and complex). The highest rate of don't knows was found on the nonsensitive complex question, and the lowest rate was found on the sensitive and complex questions. Sensitivity and complexity together seem to reduce don't know responses, while complexity in the absence of sensitivity seems to increase them. The interpretation of this result within the framework of affect and cognitive difficulty is not clear. With only one sensitive, complex and one nonsensitive, complex item it is difficult to discount the effect of individual questions beyond their rated sensitivity and complexity.

Figure 18: Interaction of Sensitivity and Complexity on Explicit Don't Knows



Another unique pattern was found for the interactions of sensitivity and complexity on answering (with and without a qualification). Both of these behaviors only showed differences between levels of complexity when questions were also sensitive, and no differences when questions were nonsensitive. There were more answers with and without qualification on the sensitive and complex question (relative to the sensitive, noncomplex question). The sensitive and complex question also had the highest rates of answering over all. While both of these behaviors cannot be applied to one utterance, both codes can be applied multiple times within the question. While a paradigmatic response would only include one answer without qualification, respondents may be asked to repeat or refine their answer, leading to a higher proportion of utterances within the question in which the respondent is answering (with or without qualification). Respondents could also modify their answers without interviewer prodding, for example answering with qualification, and then settling on a final answer (one conversational turn, but two coded utterances). This would also increase the rate of answering. Whatever the particular pattern, it is interesting that these indicators are most prevalent on questions that are both sensitive and complex, suggesting that there is something about these question demands (or perhaps this particular question) that require more attempts to answer. In the analyses above it was suggested that higher rates of answering could be due to an increase in answering (the numerator of the proportion) or a decrease in the number of utterances at the question (the denominator of the equation). This latter interpretation, that sensitive questions have shorter, more paradigmatic exchanges (e.g., approaching one respondent turn with one answer, or 1/1=1.0 in proportion of utterances answering), and that this effect is exaggerated for sensitive, complex questions is consistent with the hypothesis that respondents will talk less on sensitive and complex questions.

Figure 19: Interaction of Sensitivity and Complexity on Answer with No Qualification

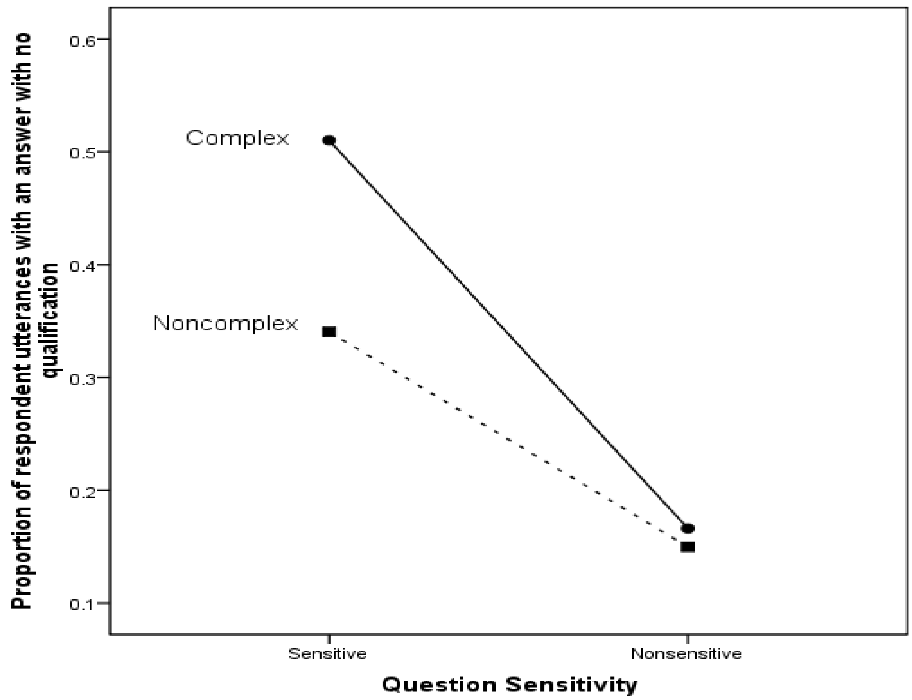
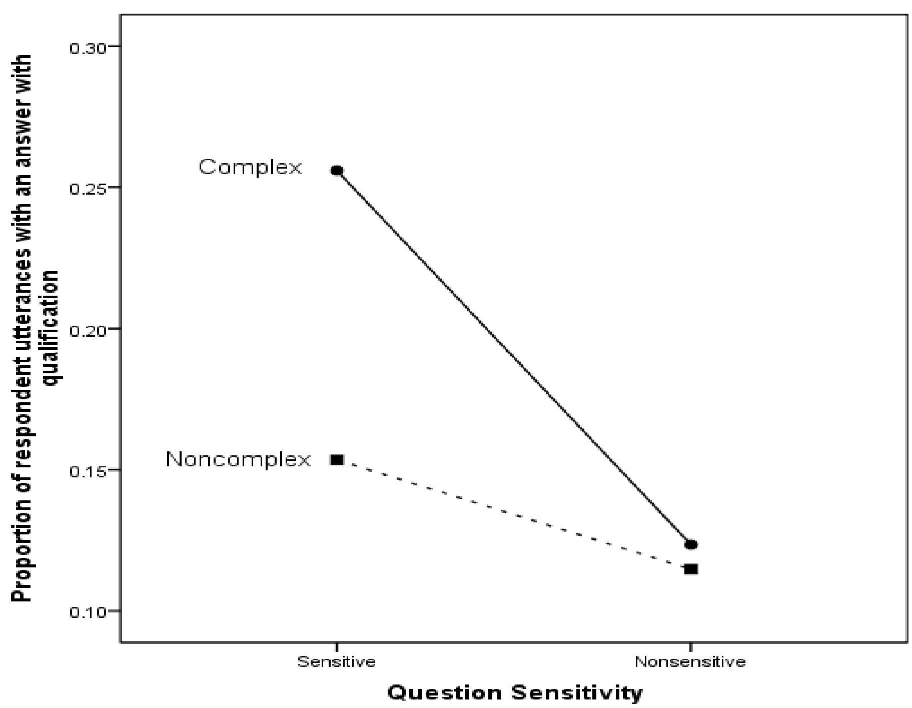


Figure 20: Interaction of Sensitivity and Complexity on Answers with Qualification



Overwhelmingly, the interaction patterns show that differences between complex and noncomplex questions were often only present when questions were also nonsensitive, with the highest rates of many affect and cognitive difficulty indicators found on the nonsensitive, noncomplex question. This pattern was found for laughter, pitch variability, and pitch range (i.e., hypothesized indicators of affect) and uncertainty about the question, filler presence and duration, pause presence and duration, respondent words spoken, stammers, repairs and stammers together, and digressions without an answer (i.e., hypothesized indicators of cognitive difficulty).

Hypothesized relationships between question characteristics were largely not supported by the interactions between sensitivity and complexity. If question sensitivity or complexity were causing laugher, disfluency, and digressions as hypothesized we would expect to see more of these on questions that are sensitive and complex. Rather we tend to find more of these behaviors on items that are neither sensitive nor complex. An alternate explanation may fit the results better. Question characteristics may still be causing these indicators, but the relationship may be the opposite of what was predicted. Questions that are sensitive and/or complex may reduce verbal indicators of affect or difficulty. Respondents may be under such intense question demands that all available cognitive resources are used to think about how and whether to answer. Further, they may choose not to display any conversational cues that encourage more discussion on the topic. When questions are neither sensitive nor complex, respondents may be able to devote more psychological resources to interacting with the interviewer through conscious effort or unconscious reactions.

3.6 Summary of Question Sensitivity and Complexity Effects

The effects of question sensitivity and complexity on respondent verbal paradata were largely counter to hypotheses or were not hypothesized at all. It was hypothesized that question sensitivity would *increase* indicators of affect (not cognitive difficulty), and that question complexity would *increase* indicators of cognitive difficulty (not affect). These hypotheses were based on an "affect/difficulty heightening" mechanism, through which question characteristics would activate the production of verbal paradata that reflect a respondent's affective or cognitive psychological state. If a respondent's subjective experience of feelings/difficulty and verbal behavior are positively correlated, and if the categorization of the four questions based on their characteristics is correct, we would expect verbal paradata to increase when questions have sensitive or complex content. This was not found. Rather than supporting an affect/difficulty heightening mechanism, the results generally support an affect/difficulty dampening mechanism. More paradata were found when question sensitivity and complexity were *absent*. Affect and difficulty expression seems to be suppressed in the presence of item sensitivity and complexity. While the main effects of sensitivity and complexity should be interpreted with caution where interactions are present, interactions largely support the same general conclusion, that is, question demands produce less, not more, verbal paradata.

Whether respondents' true affect and difficulty are heightened is unknowable in this study. Lower rates of paradata in the presence of question demands need not suggest that respondents' affect and difficulty are actually reduced. The effect of question characteristics on respondents' psychological states may be as predicted, but the relationship of those states to behavior, speech, and voice may be opposite of what was predicted. Question sensitivity and complexity may indeed cause heightened affect and cognitive difficulty, but those psychological states may result in less verbal paradata rather than more. For example, heightened anxiety may lead to an attempt to finish the survey questions faster, reducing conversation about an uncomfortable topic, thus reducing anxiety. Alternatively, and probably more likely, heightened anxiety leads to reduced psychological resources, included resources used to communicate. We see lower rates of conversational behavior on sensitive questions that support such an interpretation. The same mechanism seems to be behind responses to cognitively complex questions. Respondents produce less verbal paradata, perhaps intentionally, which likely also shortens interactions.

The findings may have implications for theory about respondent verbal behavior in survey interviews, and how we classify and use paradata that can be gathered from recordings of interviewer-respondent interactions. Specifically, they may be applicable to the relationship between verbal paradata and respondents' psychological states, which may not be as straightforward as predicted. However, inference about potential mechanisms leading to more or less respondent verbal paradata should be tempered by the reality of the study design. The design has only four questions. One of the two questions that are sensitive is also complex (the other is not complex), and one of the questions that is not sensitive is also complex (the other is not complex). Said another way, there is only one question that is neither sensitive nor complex, the condition for which various paradata were found to take up the largest proportion of utterances in many cases. Having such few questions in each condition leaves the possibility that unique characteristics of any individual question can influence the means and proportions that are being attributed to question sensitivity and complexity. They may not be representative of sensitive and complex questions more broadly. They likely also vary on other characteristics that affect verbal paradata, and this could lead to misinterpretation of the true effect of sensitivity and complexity. For example, some questions refer to the respondent, while others ask the respondent to speculate about the economy. Some questions have an open-ended format, while others do not. Table 8 outlines some of the characteristics on which these questions differ in addition to sensitivity and complexity. Some characteristics are collinear with rated sensitivity or complexity (e.g., whether the question asks about income), and so would not be separable from sensitivity and complexity in this data set. Further, none of the dimensions in Table 8 are clearly balanced (with two levels of each, e.g., single v. multiple questions), other than those that are collinear with sensitivity or complexity.

One additional question feature, question number, was explored through a repeated measures ANOVA in which question number was the only within-subjects factor and had four levels (one for each of the questions). By definition the analysis includes all within-subjects variability in one factor (question number), whereas the same variability in the sensitivity and complexity model is partitioned into two main effects terms and their interaction. Only implied refusals have a question number effect and no effect of sensitivity, complexity or their interaction. Question 2 had no implied refusals, which may lead to the effect. This question was shorter than others, having no follow-up question. The analysis by question number was not particularly informative, compared to the analysis by sensitivity and complexity. Being able to characterize the four questions along common dimensions is helpful because it allows isolation of the effects of within-subjects factors that have clearer theoretical meaning. While question number could be taken as a proxy for time or fatigue, the effect of question number did not support a fatigue hypothesis, which would be supported by clear increases or decreases in different paradata measures indicators over time (e.g., more fillers as the respondent gets tired, or short utterance durations as the questionnaire goes on and respondent loses interest). Whether sensitivity and complexity are the most informative dimensions for classification is an area for further discussion and research.

Table 8: Comparison of Question Characteristics and Dimensions they may Represent

Item	**Order**	**Sensitivity/ Complexity**	**Question Referent is Economy v. Self**	**Single v. Multiple Questions**	**Has a Qualitative Component**	**Requires Mathematical Calculation and Numeric Response**	**Asks About Income**
What about the outlook for prices over the next 5 to 10 years? Do you think prices will be higher, about the same, or lower, 5 to 10 years from now? Do you mean that prices will go up at the same rate as now, or that prices in general will not go up during the next 5 to 10 years? By about what percent per year do you expect prices to go (up/down) on the average, during the next 5 to 10 years?; How many cents on the dollar per year do you expect prices to go (up/down) on the average, during the next 5 to 10 years?	1	**Not** Sensitive, Complex	Economy	Multiple (2-3)	No	Yes (in follow-up) Estimation of %	No
During the next year or two, do you expect that your (family) income will go up more than prices will go up, about the same, or less than prices will go up?	2	Sensitive, Complex	Self (or Self and Economy)	Single	No	No	Yes
During the next 12 months, do you expect your (family) income to be higher or lower than during the past year? By about what percent do you expect your (family) income to (increase/decrease) during the next 12 months?	3	Sensitive, **Not** Complex	Self	Multiple (2)	No	Yes (in follow-up), Estimation of %	Yes
Speaking now of the automobile market - do you think the next 12 months or so will be a good time or a bad time to buy a vehicle, such as a car, pickup, van, or sport utility vehicle? Why do you say so? Are there any other reasons?	4	**Not** Sensitive, **Not** Complex	Economy	Multiple (2)	Yes	No	No

This chapter has shown that question sensitivity and complexity affect how respondents answer questions, but not as predicted. A likely mechanism behind these responses is one that causes respondents to be less conversational, whatever the cause, when question demands are high, and thus produce less verbal paradata. The following chapter will explore whether rates of verbal paradata differ between income nonrespondents, bracketed respondents, and dollar amount respondents, and whether earlier verbal paradata can predict later income nonresponse.

Chapter 4

Respondent Behavior, Speech, Voice, and Income Item Nonresponse

Analyses relating behavior, speech, and voice (i.e., respondent verbal paradata) to income nonresponse are examined in three ways. First, differences between income nonrespondent types (income nonrespondents, bracket respondents, and dollar amount respondents) on individual indicators that occur before the income question are evaluated one indicator at a time. Second, individual indicators from all questions (including income those measured at the income question) are used to predict income nonrespondent type. Finally, income nonresponse is predicted by factor scores from a basic one-factor model using indicators at each question based on the most highly correlated indicators. The goals of the chapter are to explore the relationship between income nonresponse and respondents' verbal paradata, to see if verbal paradata prior to the income question can predict income nonresponse, and to explore common factors that can explain verbal paradata.

4.1 Income Nonrespondent Type

Income nonresponse is a complex topic, and the full range of its causes is not well understood. Respondents' decisions to provide income likely result from multiple factors that act consciously and unconsciously. Some of these factors are in the control of the researcher, including the form of the question, mode of survey administration, interviewer characteristics or characteristics of the sponsoring agency. Other factors are characteristics of the respondents, and not under control of the researchers. Older respondents, White respondents, and respondents who are self-employed or employed less than full-time tend to have more income nonresponse than their younger, non-White, and fully-employed and overseen counterparts (Bell, 1984; Nicholetti & Peracchi, 2001; Riphahn and Serfling, 2005).[15] Beyond demographics, psychological characteristics (e.g., attitudes, beliefs, personality, cognitive ability, and feelings) can impact the decision about whether to provide one's income when it is requested by an interviewer. These psychological factors can be described as having both predetermined and alterable components. Predetermined components are propensities for nonresponse that are present in a respondent before the survey even begins, *and* are uninfluenced by facets of the design or situational details of the interview. Alterable components are those propensities for nonresponse that are changeable by facets of the design. For example, cognitive ability is an example of one factor that has been shown to affect income nonresponse (Juster and Smith, 1997; Heeringa, Hill, and Howell, 1993). Each respondent has a baseline cognitive ability level before they begin the survey. Some respondents will be high in cognitive ability and some will be low. Some respondents' baseline cognitive ability will completely determine their income nonresponse status. It is these cases that have a predetermined (i.e., fixed) income nonresponse propensity. For example, respondents with the lowest cognitive ability may have a propensity for income

[15] Looking at item nonresponse more generally, women, respondents with less education, and respondents in blue-collar jobs tend to have more item nonresponse than men, respondents with more education, and respondents in white-collar jobs (Craig and McCaann, 1978; Ferber, 1966).

nonresponse fixed at 1.0. No matter how simple the income question is, or how well the interviewer probes, they will not be able to answer. Respondents with moderate cognitive ability may be challenged by cognitively complex questions like income, but still be able to answer. For respondents with a propensity somewhere between 0 and 1, their initial risk of nonresponse can interact with question characteristics, interviewer characteristics, or other facets of the survey design. The technique of offering income brackets to dollar amount nonrespondents works on this principle by modifying the income question to be easier for those who initially find it difficult. Of course, bracketing does not work for every respondent, and thus in any given instance of bracket refusal an unanswerable question is whether a different modification would have obtained a response from that respondent or whether their nonresponse propensity was truly fixed at 1.

Privacy concerns are another component, likely affective in nature, that can influence income nonresponse. Some respondents will have a predetermined and inflexible attitude about not sharing income information in any circumstance, and their behavior will consistently reflect that attitude. The staunchest of the high privacy respondents are unlikely to be swayed by any modification to the survey question or statement of confidentiality. Those respondents could be said to have a fixed propensity to not provide income. Other respondents with a less extreme nondisclosure attitude might be willing to provide income with assurances of confidentiality, or to an interviewer that they deem trustworthy. These respondents would be expressing an alterable (i.e., variable) component of the privacy concerns factor. At the other end of the privacy concern continuum, someone with no concern for privacy might have a high baseline propensity for disclosing income, but decide not to do so if there is something about the question, interviewer, or sponsor that they find untrustworthy. Respondents with no privacy concerns at all (e.g., fixed propensity to respond) will be insensitive to any of these facets and always provide income data.

In summary, cognitive ability (and other attributes related to cognitive difficulty) or privacy concerns (and other attributes related to affect) are likely to be completely collinear with income nonresponse for *some* respondents. For respondents fitting this profile, relevant characteristics at the beginning of the survey are a good proxy measure for income nonresponse. Income nonresponse status is something they carry with them from before the survey even begins. For other respondents, the two statuses (cognitive ability and income nonresponse) are not so closely linked, with characteristics of the mode, question, or interviewer acting as moderating factors. These latter respondents are the ones that we can actually influence through survey design features. Yet, understanding the psychological roots of income nonresponse requires knowing something about respondents whose responses can be changed as well as those whose cannot.

Exploring the psychology of income nonresponse motivates a dual focus on respondent psychology and survey design. It also motivates the idea just forwarded that income nonresponse propensity can be thought of as a property of individual respondents. Although income nonresponse status is known only after an income question is asked, it is clear that it has predetermined components that are defined by psychological facets outside of the survey design itself. For respondents with a completely predetermined propensity, their eventual income nonresponse status can be used as a grouping characteristic, similar to age, sex, or personality characteristics. Not all respondents have such a fixed propensity. However, once known, income nonresponse status is likely closely reflective of true income nonresponse propensities that characterize respondents before an income question is asked. Nonrespondent type, once known, can be thought of as a categorization of true income nonresponse propensity, which is continuous. Thus, we can review differences in behavior, speech, and voice between income nonresponse types (i.e., nonrespondent, bracketed respondent, and dollar amount respondent). We will do this using only data that occur before the income question. This will test whether individuals who end up having trouble reporting income differ from those who report an income amount even before they hear the income question.

In this dissertation, the way respondents answer (or don't answer) the household income question puts them in one of three categories. Complete income nonrespondents provide no information about their income. This could be by explicitly refusing to answer, or by reporting that they don't know their income, and then refusing or not being able to answer in bracket form. Bracket respondents are respondents who, for some reason, cannot or will not provide a dollar amount, but give an answer to a series of income values that assign their income to a range (e.g., above $50,000, but below $60,000; see Appendix J for the bracketing procedure used by the SCA). Dollar amount respondents provide a specific income value to the question about household income in the past year. This three-level classification of income nonresponse will be referred to in the rest of the dissertation as "income nonrespondent type" or just "nonrespondent type". Each respondent is in one and only one of these categories based on the final result of the income question. Income nonrespondent type was known when the cases were selected from the Surveys of Consumers, and was corrected if the original classification seemed to be in error after listening to the recording.

4.2 Income Nonrespondent Type and Respondent Behavior, Speech, and Voice

Respondent behavior, speech, and voice on four questions before income define a series of repeated measures that were analyzed in Chapter 3. As in Chapter 3, none of the measures use in this analysis occur on the income question itself, thus significant effects of income nonrespondent type should be interpreted as differences in respondents who *eventually* answer the income question in a particular way (e.g., refusal/don't know, brackets, or dollar amount). Significant main effects of income nonrespondent type on individual indicators are evaluated first through a one-way ANOVA that averages indicators over the four questions before income. Table 9 includes only those indicators for which there is a significant effect of nonrespondent type. The full list of indicators that were used as dependent variables in this analysis can be seen in Appendix G. The table presents each indicator, the F-value and associated p-value for the significance of the model, the direction of the difference between nonrespondent types, associated p-values for post-hoc comparisons, and means for each nonrespondent type group (NR=nonrespondents, BR=bracketed respondents, DA=dollar amount respondents)[16]. The one-way ANOVA for each indicator defines income nonrespondent type as the independent variable and the respective indicator as the dependent variable.[17] An α-level

---

[16] As in Chapter3, the means reported are based on question-level averages and proportions of utterances per question so that question length does not artificially afffect the presence of indicators.

[17] For ease of interpretation, and to meet the goals of this chapter, only the one-way ANOVA results are reported here. More complex models were tested as well, including general linear models that included

of .05 was used to determine what to present in the table. Full ANOVA results, F-values, and p-values are presented in Appendix K.

A picture emerges showing that income nonrespondents and bracketed respondents are both identifiable by respondent verbal paradata that occur before the income question. Respondents who eventually became income nonrespondents had more negative comments than bracketed respondents or dollar amount respondents. It is intuitive that that respondents who have something negative to say about the survey would also be less likely to provide income information, as a negative comment could reflect distrust, frustration, or general displeasure with the survey experience. Affect intensity and affect valence also showed that respondents who eventually became income nonrespondents had lower intensity and negative rated affect on average. Both dollar amount respondents and bracket respondents had higher and positive rated affect compared to income nonrespondents.

Income nonrespondents also differentiate themselves on the rate of digressions with no codable answer. Respondents who end up becoming income nonrespondents seem to get off track without providing content that is codable by the interviewer (i.e., content related to the question) more often than bracketed respondents and income nonrespondents. This was the only cognitive difficulty indicator that distinguished income nonrespondents from the other two nonrespondent types.

---

sensitivity and complexity as within-subjects factors and all interactions between predictors, and only main effects for sensitivity, complexity, and nonrespondent type. The nonrespondent type main effects are robust and show up in identical form in all models tested. Exploration of interactions between nonrespondent type and question characteristics will be the topic of later research on these data.

Table 9: Summary of Significant Differences in Nonrespondent Types (α=.05 level only)

Construct	**Indicator**	**Direction of Difference**	**Mean Proportion of Utterances with Indicator per Question**
Affect	Negative comments *($F^{\dagger}$=5.614, p=.004)*	More in income nonrespondents than bracket respondents *(p=.008)* More in income nonrespondents than dollar amount respondents *(p=.018)*	NR*: .006 BR: 1.5E-019 DA: .001
	Affect intensity *(F=6.67, p=.002)*	Lower in income nonrespondents than bracket respondents *(p=.001)*	NR: 4.238 BR: 5.987 DA: 5.178
	Affect valence *(F=8.825, p<.0005)*	Lower (more negative) in income nonrespondents than bracket respondents *(p.041)* Lower in income nonrespondents than dollar amount respondents *(p<.0005)*	NR: -.058 BR: .182 DA: .325
Cognitive Difficulty	Digression with no answer *(F=6.042, p=.003)*	More in income nonrespondents than bracketed respondents *(p=.010)* and dollar amount respondents *(p=.007)*	NR: .092 BR: .046 DA: .046
	Digression with a codable answer *(F=5.128, p=.007)*	More in bracket respondents than income nonrespondents *(p=.005)*	NR: .013 BR: .049 DA: .033
	Report *(F=2.878, p=.059)*	Borderline more in bracket respondents than dollar amount respondents *(p=.053)*	NR: .113 BR: .148 DA: .090
	Cognitive difficulty rating[18] *(F=34.427, p<.0005)*	More in bracket respondents than nonrespondents *(p<.0005)* More in bracket respondents than dollar amount respondents *(p<.0005)*	NR: .102 BR: .425 DA: .169

*NR=Nonrespondent, BR=Bracket respondent, DA = Dollar amount; †*F(2, 182)* for all F-tests

[18] The same pattern holds when each category of this three-category variable is analyzed independently. For the "no cognitive difficulty" category, more was found in nonrespondents and dollar amount respondents than bracketed respondents *(F=34.102, p<0005)*. For the "some cognitive difficulty category", more was found in bracket respondents than income nonrespondents and dollar amount respondents *(F=34.427, p<.0005).* For the "high difficulty" category, more was found in bracket respondents than income nonrespondents only *(F=3.058, p=.049).*

Respondents who eventually answered income with brackets were identified by more digressions with a codable answer, more reports, and higher rated cognitive difficulty than nonrespondents and dollar amount respondents on survey questions before the income question. Bracketed respondents seem to get off track while also providing an answer more than income nonrespondents. They also offer answer-relevant information without answering the question (i.e., report) more often than income nonrespondents and dollar amount respondents, but this effect was borderline at the α=.05 level. These behaviors combined with higher cognitive difficulty ratings are consistent with the hypothesis that bracketing relieves cognitive burden in income reporting (Juster and Smith, 1997; Heeringa, Hill, and Howell, 1993), and also suggest that the cognitive burden relieved by bracketing may not be caused only by difficulty reporting income. Rather, it may also be present on items before the income question, and perhaps due to respondents' general cognitive difficulty with survey questions or general cognitive ability.

Summarizing the differences in indicators between nonrespondent types it is clear that respondents who do not provide full income information (e.g., income nonrespondents and bracketed respondents) distinguish themselves on paradata that they produce prior to the income question. Nonrespondents were identifiable primarily on affect indicators (negative comments, valence, and intensity) and one cognitive difficulty indicator (digressions without a codable answer). Bracketed respondents were identifiable only on cognitive difficulty indicators (digressions with an answer, reports, and rated cognitive difficulty). These findings support previous research showing that brackets alleviate cognitive difficulty in reporting income, but extend that explanation by showing that difficulty leading to the choice of brackets may also be seen on questions prior to income. The findings also suggest that income nonrespondents may not be providing income due to affective reasons more than cognitive reasons. Income nonrespondents seem to have more negative feelings about the survey experience in general, evidenced by more negative comments and more negative rated affect than other income nonrespondent types. Digressing without providing a codable answer was the only indicator of cognitive difficulty on which nonrespondents showed the highest rates.

The ANOVA model used places some constraints on the inference that can be drawn about the cause of income nonresponse. The statistical model used predicts indicators from income nonrespondent type. To meet the goals of the broader project, income nonresponse should be on the left side of the equation that is used to model the data, not on the right as it is with the current analysis. The direction of prediction will be reversed in the next section. As income nonrespondent type is known only after the income question is asked (i.e., at the end of the survey for our purposes), it may be counter-intuitive to think of this status as a right-side factor (i.e., “causing” the indicators if the equation is taken causally). The interpretation of results under the current model is aided by assuming that a respondent’s income nonresponse status is fixed prior to the income question. Income nonrespondent type is present but unknown until the income question is asked. If we accept this assumption, we can argue that respondents’ unknown-but-present income nonrespondent status causes indicators that occur before the income question. For certain respondents, this may in fact be the case. Respondents who never provide their income (or who always provide their income) may come into the interview with a predetermined and fixed income nonresponse status. If they are strong in their conviction, and thus uninfluenced by essential survey conditions or interventions, then this status can be considered as a between subjects factor equivalent to sex, age, or education. Understanding the true flexibility of this income nonrespondent status would require a repeated measures design in which respondents answer income questions that vary on essential survey conditions (e.g., mode, interviewer characteristics, form of the question, sponsor), testing which specific features can turn a nonrespondent into a respondent or vice versa. This is beyond the goal of this dissertation. The next section will evaluate the ability of individual indicators and factor scores made by combining indicators to predict income nonrespondent type.

4.3 Affect and Cognitive Difficulty Indicators, and Factors that Predict Income Nonrespondent Type

The analytic goal of this section is to predict income nonrespondent type. The previous analysis documents differences between income nonrespondent types, where nonrespondent type was an independent variable in the analysis (i.e., a predictor of individual indicators). The analyses presented here use nonrespondent type as an outcome. Predictors consist of individual indicators in the first analysis and factor scores derived from these indicators in the second analysis.

4.3.1 Individual Indicators Predicting Nonrespondent Type

The previous analysis explored whether income nonrespondent types would show different rates of individual indicators, and was done one indicator at a time. The current analysis uses the same set of indicators (with minor recoding) to predict income nonrespondent type. This analysis has two benefits over the previous one. First it avoids the need to assume that income nonrespondent type is a fixed characteristic of respondents. The goal here is to predict nonrespondent status from indicators that arise in respondent actions before and during the income question. Second, it puts indicators in direct competition with each other, with the goal of finding which indicators best predict income nonrespondent type when all indicators are analyzed together.

The data source for these analyses is essentially the same as that used in Chapter 3, but with some restructuring of the data and recoding of individual variables. For the analysis using individual indicators as predictors, the data file was restructured from a wide format (one case per respondent with repeated measures for each question) to a long format (five cases per respondent, one for each question).The CLUSTER command was used in Mplus to define respondents as clusters and adjust p-values and standard errors for the clustering of observations at each question within respondents.

During factor analysis, some indicators were recoded to simplify the models and improve the chances of convergence and fit. Recoding was based on the skew of the raw variables, most of which had zero-inflated distributions. Also, indicators that we distinguished in coding (e.g., implicit and explicit don’t knows) were combined with each other. Specifically explicit and implicit don’t know and refusal codes were recoded into one don’t know code and one refusal code. Repairs and stammers were recoded into one repair or stammer variable. Expressions of uncertainty, whether about the question or how to answer, were combined into one variable. All of these indicators were re-coded to be binary at the question level. Presence at the question, regardless of frequency, was coded as 1 and absence was coded as 0. The following variables were also recoded such that any occurrence received a score of 1 and no occurrence received a score of 0; backchannels, conversation management, laughter, negative comments, answers without qualification, answers with qualification, clarification and repeat requests, reports, fillers, and pauses. The variables analyzed here, and their recoded distributions are in Table 10.

Table 10: Indicators Recoded from Continuous to Binary Variables

Recoded Indicator	**Proportion of Occurrence ≥ 1 (all 5 questions)**
Hypothesized Indicators of Affect	
Refusal (explicit and implicit)	.125
Negative comment about the survey	.014
Laughs	.151
Backchanneling	.093
Conversation management	.068
Hypothesized Indicators of Cognitive Difficulty	
Answering primary question	.528
Answering primary question with qualification	.366
Request for clarification or repeat of question	.169
Uncertainty about the question or their answer	.117
"Don't Know" (explicit or implicit)	.154
Digression (with and without codable answer)	.239
Report	.258
Repair or stammer	.420
No difficulty	.930
Filler	.467
Pause	.277

Table 10 does not include all the predictors used in this analysis. Indicators that were less skewed or did not make sense as a binary variable (e.g., pitch) were entered in their continuous form. These variables and their means were affect intensity (*Mean*=5.84), affect valence (*Mean*=.071), median pitch (*Mean*=158.19), pitch 5th percentile (*Mean*=119.8), pitch 95th percentile (*Mean*=264.98), pitch standard deviation (*Mean*=50.65), pitch in last 50ms of voicing (*Mean*=183.06), speech rate in syllables per second (*Mean*=4.7), speech rate in words per second (*Mean*=1.06), length of exchange in utterances (*Mean*=9.57), duration of first respondent utterance (*Mean*=3.041), average words spoken per question (*Mean*=105.4), and average overspeech per question (*Mean*=3.051).

Table 11 shows the significant coefficients from a forced multinomial logistic regression model with all indicators included (i.e., no model selection)[19]. All coefficients from the full model are reported in Appendix L. The regression results show that far more indicators predict income nonresponse (relative to dollar amount response) than predict bracketed response (relative to dollar amount response). Income nonresponse is predicted by more negative comments, reports and refusals, less backchanneling, question-answering (with or without qualification), and lower ratings of affect intensity and affect valence relative to dollar amount respondents. Bracketed response was only predicted by more reports and more refusals relative to dollar amount respondents on all questions including income.[20]

[19] The multinomial logistic regression produces generalized logits. The coefficients can be interpreted as reflecting the change in the logit of the relevant income nonrespondent type category relative to dollar amount respondents for a one unit change in the predictor, with all other predictors in the model and held constant. The coefficients are all unstandardized, as standardized coefficients are not available Mplus for categorical predictors. As with previous analyses, an α-level of .05 was used to determine what to report in the table.

[20] Other than the recodes discussed, the variables that were analyzed individually in the one-way ANOVA were used in this analysis. Additionally, pitch range was excluded due to its colinearity with other pitch measures. When included, Mplus fixed estimates for this variable, so it seemed more parsimonious to exclude it.

Table 11: Significant Predictors of Income Nonresponse (α=.05)

Indicator	Estimate	S.E.	p-value
Income Nonresponse			
Backchannels	-0.902	0.385	0.019
Affect Intensity	-0.207	0.055	< 0.0005
Affect Valence	-0.713	0.160	< 0.0005
Overspeech	0.069	0.025	0.006
Negative Comments	2.081	1.107	0.06
Answering without Qualification	-1.109	0.298	< 0.0005
Answering with Qualification	-0.849	0.285	0.003
Reports	0.735	0.286	0.01
Refusals	1.472	0.474	0.002
Don’t know	0.657	0.298	0.027
Bracketed Income Response			
Reports	0.714	0.275	0.01
Refusals	1.685	0.476	< 0.0005
No difficulty	-1.262	0.468	0.007

The log odds of income nonresponse relative to dollar amount response are predicted by *less* backchanneling, *lower* affect intensity, and *less positive* (on average negative) affect valence. In other words, income nonresponse is strongly predicted by less communicative behavior and affective involvement in the survey interaction than dollar amount response. Yet, in other ways, income nonresponse is predicted by more conversational behavior than dollar amount response, such as overspeech, reports, and the amount of negative commentary (marginally). If respondents are talking more, there are more opportunities for interviewers and respondents to talk over each other (overspeech). From these data, however, it is not known whether the overspeech involves respondents interrupting interviewers, interviewers interrupting respondents, or whether it is overspeech with no interruption. It is also not known whether the interruption seems to hinder the conversation, or whether it is dealt with, and the conversation moves forward. These are two qualitatively different examples of how overspeech could occur.

The expression of negative comments could also be considered a conversational behavior, in that respondents who make these comments are speaking more frankly with the interviewer than those who do not. Although the expression of negative comments is conversational in one respect, it does not seem like a conversational behavior that encourages more interaction.

Reports are another conversational behavior that positively predict the log odds of income nonresponse (i.e., responses that contain information about their answer but do not answer the survey question directly). Although we do not know the content of or motivation for the reports (e.g., confusion, embarrassment), reports involve talking more, rather than talking less. Reports also indicate the respondent’s effort toward providing an answer the interviewer could accept. On one hand, they represent incomplete answers, but on the other hand they represent attempts at providing a complete answer. Regardless of the motivation, they involve increased verbalization.

Not surprisingly, higher rates of refusals and don’t knows, and lower rates of question-answering (with and without qualification) predict respondents who do not report income. Refusals seem to predict income nonresponse more strongly than don’t knows, but it is interesting that both predict the log odds of not reporting income. This confirms the assertion that income nonresponse is related to not *wanting to* respond for some respondents (e.g., refusals), and not *being able to* respond for other respondents. From this analysis it appears that refusals might predict income nonresponse more strongly than don’t knows, supporting a motivational or affective explanation, with difficulty responding as a secondary explanation.

These findings partially replicate what was found in the income nonrespondent type analysis, where negative comments were higher in income nonrespondents than the other two nonrespondent types, and affect intensity and valence were lower. In that analysis, reports were found more frequently in income nonrespondents than dollar amount respondents, but the difference was not significant at the α=.05 level. No income nonrespondent type differences were found in the first analysis for backchannels, answering (with or without qualification), or refusals.

Bracketed response was only predicted by *more* refusals and reports relative to dollar amount respondents, and *more* difficulty (e.g., less questions with a code of "no difficulty). It is interesting that two of the indicators that increase the log odds of bracketed response also predict income nonresponse (refusals and reports). The ANOVA results found that reports (borderline at α=.05), cognitive difficulty, and digressions with a codable answer were found at higher rates in bracketed respondents compared to the other income nonrespondent types. Both analyses agree that reports are predictive of bracketed response, but differ on other variables. As was seen in the ANOVA results, more rated cognitive difficulty predicts the log odds of becoming a bracketed respondent, relative to a dollar amount respondent. This variable also uniquely predicts bracketed response, and does not predict income nonresponse. Supporting the ANOVA results, it seems that respondents who have problems with survey questions in general will likely become bracketed respondents on income questions.

The differences and similarities in significant predictors of income nonresponse and bracketed response warrant further discussion. At the individual level, a person with a high rate of refusals and reports is likely expressing difficulty or discomfort with the survey questions, and should be more likely to not report income, or to only report it in bracketed form. However, if they also exhibit low levels of backchanneling, lower perceived affect intensity and valence, less question-answering, more refusals, and perhaps most saliently, more negative comments, they are clearly expressing some sort of discomfort, irritation, or affective disengagement that will lead to income nonresponse. If they express more difficulty, they are likely going to report their income in brackets.

Reports and refusals seem to play a unique role in understanding the mechanisms behind income nonresponse and bracketed income response. Both predict problems reporting income. It is theoretically interesting that refusals, but don't know responses predict bracketed response. Don't know responses would be expected to predict bracketed response if they are indeed an indicator of cognitive difficulty answering survey questions. Yet in this analysis, they predict income nonresponse and not bracketed response. Refusals, however, do predict bracketed response.

Like refusals, reports predict difficulty with income reporting generally. While reports are often tied to cognitive difficulty (Maynard & Schaeffer, 2002a, Schaeffer & Maynard, 2007; Schober & Bloom, 2004) they can also indicate discomfort with a sensitive topic. In this coding scheme, the only instruction for coding a report was that the respondent gives answer-relevant information without answering the question. Examples of a report due to cognitive difficulty on an income question might be "I work two jobs so it's complicated" or "I made forty-thousand for half the year, and then I got a raise and now I make fifty-thousand". Compare this to reports that are more likely affective in motivation, "Things are tight these days because my wife just lost her job" or "We do alright". Reports are simply behaviors, and the cognitive or affective motivations behind them may not always be clear. From the positive prediction of both refusals and reports, we might expect that these reports are more affective than cognitive in content.[21]

This analysis showed that income nonrespondents display many signs that they may refuse to provide income. Bracket respondents provide only three indicators, and two of those are shared with income nonrespondents. The motivation for complete absence of income information seems to be more affective, while the motivation for bracketed response seems more cognitive. Yet it is difficult to interpret causal mechanisms when indicators are reviewed individually and results are not overwhelmingly in one direction.

4.3.2 Factors Predicting Income Nonrespondent Type

The best way to model the psychological factors (and likely mechanisms) behind income nonresponse would be through a full latent variable model or structural equation model (SEM). This requires a well-fitting measurement model (factor structure) composed of the indicators that have thus far been analyzed independently. Such a model was attempted through exploratory and confirmatory factor analyses, but no model was found that fit the data well by accepted measures of fit (e.g., RMSEA).[22] To reduce the

---

[21] However, the bivariate correlations show that don't knows are highly correlated with reports (r=.423) while refusals are not.

[22] The factor analytic steps began with a confirmatory factor analysis based on the predicted relationships of each indicator to affect and cognitive difficulty factors. Modifications were made based on cross-loadings in this initial hypothesized model. Neither model fit the data well. The bivariate correlations between indicators were reviewed via a correlogram (see Appendix M) and additional confirmatory factor models were fit, but these did not meet accepted fit criteria either, likely due to the lack of clear factors as evident in the correlogram. An exploratory factor analysis was conducted as well, but models tested either did not converge or presented ambiguous factor solutions. The difficulty in finding a factor structure with reasonable fit lead to the decisions analyze factor scores directly as a data reduction technique.

number of predictors and provide a more parsimonious model predicting income nonrespondent type, factor scores were calculated, exported, and used as predictors in the multinomial regression model presented here.[23] To address the temporal sequencing of respondent behavior across questions in the survey, factor scores were exported at each of five questions, providing the ability to test whether a single simple factor predicts income nonresponse differently at each of the five questions. One explicit question of this dissertation is whether respondent verbal paradata before the income question predict income nonresponse better or worse than paradata at the income question. This will be addressed here.

Factor scores were based on a one-factor model that uses the most highly correlated indicators (see Appendix M for the correlogram of indicators). These indicators are the variables most likely to be part of the true factor structure if a factor model were developed that fit the data well. Due to high bivariate correlations that cross predicted factor boundaries, an intensive modeling of cross-loadings would be required to develop a model that fit the data well. The approach used here subsumes all factor loadings (including cross-loadings) and measurement error into one factor score made up of the following variables: affect intensity, laughter, pitch (95^{th} percentile, standard deviation, and median in last 50 ms of voicing), refusals, number of utterances, digressions (with and without qualification), repairs and stammers, answering (with and without qualification).

[23] This approach does not fully model the relationships between indicators, and the one factor specified does not fit the data well. This analysis sacrifices measurement exactness for data reduction and model simplicity.

For the factor score export, cases were selected from each question individually, and a repeated measure (wide) dataset was build with one vector of factors at each question. The distributions of the factor scores at each question are as follows; Question 1 (nonsensitive, complex; *Mean=.015, SD=.42),* Question 2 (sensitive, complex; *Mean=0.0, SD=1.0),* Question 3 (sensitive, noncomplex; *Mean=.011, SD=.64),* Question 4 (nonsensitive, noncomplex; *Mean= .005, SD=.48),* Question 5 (income; *Mean=0.0 SD=1.16).*

When one factor score per question (e.g., five predictors) was entered into a multinomial logistic regression analysis, results show differences between predictive ability at each question. Using an alpha level of .05, factor scores on the nonsensitive, noncomplex question (Question 4) are the only scores that positively predict the log odds of income nonresponse relative to dollar amount response. High values of the Question 4 factor predict a lower log odds of becoming an income nonrespondent. Questions 1-3 seem to have no effect on the log odds of becoming a nonrespondent. It may be the case that income nonrespondents are so inherently different from other respondents that they act differently on a question that is neither sensitive nor complex (Question 4). Every other question is either sensitive or complex, or both. Question 4 is neither. It is interesting that a question that should provide no particular challenge produces paradata that significantly predicts income nonresponse. This appears to support conclusions developed above that those who do not report income telegraph their intensions through verbal paradata before the question. The income question's (Question 5's) factor scores approach significance ($p=.061$). If we interpret this at the more liberal $\alpha=.1$ level, we can conclude that paradata at the income question also predicts income nonresponse, which should be expected. Table 12 summarizes the coefficients for each factor score predicting income nonresponse and bracketed response.

Table 12: Multinomial Regression Coefficients for Factor Scores Predicting Nonrespondent Type

Factor	Estimate	S.E.	p-value
Income Nonresponse			
Question 1 Factor Scores (Nonsensitive, Complex)	0.245	0.413	0.553
Question 2 Factor Scores (Sensitive, Complex)	0.102	0.199	0.609
Question 3 Factor Scores (Sensitive, Noncomplex)	-0.122	0.313	0.698
Question 4 Factor Scores (Nonsensitive, Noncomplex)	***-1.008***	***0.313***	***0.009***
Question 5 Factor Scores (Income)	0.771	0.411	0.061
Bracketed Response			
Question 1 Factor Scores (Nonsensitive, Complex)	0.450	0.492	0.360
Question 2 Factor Scores (Sensitive, Complex)	0.085	0.212	0.687
Question 3 Factor Scores (Sensitive, Noncomplex)	-0.800	0.428	0.062
Question 4 Factor Scores (Nonsensitive, Noncomplex)	-0.555	0.459	0.227
Question 5 Factor Scores (Income)	***1.505***	***0.521***	***0.004***

The log odds of becoming a bracketed respondent relative to a dollar amount respondent are predicted only by factor scores at Question 5 (the income question), with the sensitive, noncomplex question's (Question 3) factor scores approaching significance. Again, it is not too surprising that behavior at the income question predicts incomplete income data. It is interesting that this effect is stronger and more highly significant for bracketed respondents than it is for income nonrespondents, suggesting that bracketed response may be more identifiable at the income question than it is before the income, while income nonresponse is more identifiable before the income question. If we interpret Question 3's coefficient at the more liberal α=.1 level, we see that bracketed respondents' paradata on a question high in sensitivity predicts their income nonrespondent status. This differs slightly from expectations developed from previous analyses, where it might be expected that verbal paradata on cognitively difficult questions are better predictors of bracketed response.

The results of the factor score analysis suggest that complete income nonresponse potential can be seen before the question, but only on a question that is neither sensitive nor complex. Income nonresponse might also be predictable at the income question with a larger sample size. Bracketed response however is predicted by behavior at the income question, and only marginally by behavior before. Considering the idea that income nonresponse is partly a fixed characteristic of respondents, it could be concluded that those respondents who incognito income nonrespondents reveal their income nonrespondent type on a question that shouldn't provide a challenge for respondents. Bracketed respondents, on the other hand, likely have a more flexible income response propensity (i.e., one that is not predetermined before hearing the income question, and one that may in fact be influenced by the presentation of brackets), and thus their final income nonresponse status is best predicted by their paradata at the income question.

With only one factor specified, and without a good measurement model for that factor, the true mechanism leading to income nonresponse or bracketed response is difficult to discern. The variables used in the factor score calculation represent hypothesized affect indicators (affect intensity, laughter, pitch, and refusals) as well as indicators hypothesized to represent cognitive difficulty (number of utterances, digressions, repairs and stammers, answering with and without qualification). They include measures of question-answering behavior (refusals, digressions, and answers with and without qualification), speech (number of utterances, laughter, repairs and stammers), voice pitch, and rated affect (intensity). The most that can be concluded is that some combination of affect and cognitive difficulty indicators predict incomplete income data, and that the complete lack of income data can be predicted by a respondent's voice, speech, and question-answering before they hear the income question.

4.4 Nonrespondent Analysis Summary

The analyses presented here had the overlapping goals of exploring differences between income nonrespondent types and exploring what predicts income nonresponse. Analyses of individual indicators came to similar but not identical conclusions. When only data before the income question were analyzed, only those respondents who had difficulty with income (nonrespondents and bracketed respondents) distinguished themselves on individual indicators of affect and cognitive difficulty. Income nonrespondents gave more negative comments and exhibited lower perceived affect intensity and negative affective valence. They also produced more digressions with no codable answer, which is a hypothesized indicator of cognitive difficulty. Bracketed respondents showed more digression with a codable answer, marginally more reports, and higher perceived cognitive difficulty on questions prior to the income question. These were all hypothesized indicators of cognitive difficulty. A picture of income nonresponse begins to emerge in which total nonresponse is preceded by negative affect cues, and bracketed response is preceded by cognitive difficulty cues.

When individual indicators were used to predict income nonresponse, results look similar. Income nonresponse was still predicted by negative comments, lower affect intensity and negative affect valence. Backchanneling (less), overspeech (more), answering the survey question (less), refusals (more), and don't knows (more), also predicted income nonresponse. The image of income nonrespondents is one that includes less backchanneling, a productive communicative behavior, but more negative comments, an unproductive one. Overspeech, a neutral communicative behavior, also predicts income nonresponse. In the affective dimension, income nonresponse was also predicted by flatter affect relative to dollar amount response. Income nonrespondents are more conversational when it comes to expressing negative feelings about the survey or providing answer-relevant information without actually answering the question (reporting), relative to dollar amount respondents, but these may not be helpful of the goal of answering the survey question.

The image of bracketed respondents is also somewhat similar between the two analyses. When using data before the income question only, it was shown that respondents who have difficulty answering income questions produce more reports, more digressions (with a codable answer), and higher rated cognitive difficulty than other income nonrespondent types. This is evidence that respondents who eventually need or choose to use brackets to report income are having trouble with survey questions in general. Rather than being unique, income may be just another survey question that these respondents find difficult to answer. While the predictors for bracketed response were not identical in the multinomial logistic regression, the picture looks similar with respect to cognitive difficulty. In the regression analysis, reports, refusals, and cognitive difficulty predict bracketed response. Reports and cognitive difficulty ratings support the cognitive difficult argument, but in this analysis reports also predict complete income nonresponse. Refusals may suggest an affective component of bracketed nonresponse, counter to the cognitive difficulty one just summarized. Despite the differences, both analyses together point strongly toward cognitive difficulty related to bracketed income response. More precise conclusions about the role of reports in predicting income reporting could be drawn by looking at specific instances of reporting, and perhaps coding them for affective or cognitive content. Some reports may reflect lack of information (e.g., cognitive reasons), while others may reflect discomfort with the question (e.g., affective reasons). Such an analysis could help support this distinction between types of reports.

Analysis of factor scores, using the most highly correlated individual indicators, showed that income nonresponse was predicted by behavior, speech, and voice prior to the income question. Respondent paradata on the nonsensitive, noncomplex question predicted income nonresponse. Respondents who are having some sort of affective or cognitive reaction to the survey interview as a whole should also express behavior, speech, and voice cues on questions that are not threatening or difficult (i.e., questions that don’t cause problems for other types of respondents). The fact that behavior, speech, and voice on such a question predict income nonresponse supports this assertion.

Bracketed respondents are generally thought to have difficulty with income specifically, which is why they report in brackets rather than dollar values. The analysis of factor scores supports that idea. Factor scores at the income question predict bracketed response, while factor scores on other questions do not. At the same time, other analyses presented here show that respondents who eventually report income with brackets provide evidence of more cognitive difficulty prior to the income question. This suggests that bracketing may be a result of a general cognitive difficulty profile that is not limited to the income question alone. Yet when all verbal paradata are taken as a whole (i.e., factor scores) indicators at the income question predicted bracketed response best. While these findings may seem contradictory, they are not necessarily at odds with each other. In the individual indicators ANOVA results, no indicators that occurred on the income question behavior were used, and indicators were each analyzed individually. In the factor score analysis, income question data were used and indicators were combined together into a single factor score at each question. Indicators that clearly measure cognitive difficulty and affect, as well as less diagnostic verbal paradata are lumped together, and so it is not completely surprising to find different results between individual indicators and factor scores.

The totality of these results paints a picture of income nonresponse as a response propensity that has observable antecedents, some of which occur even before the income question is asked. Income nonrespondents produce a plethora of individual indicators that telegraph their intent to not provide income, and this finding holds up when indicators are summarized into factor scores. Bracketed respondents show fewer indicators, but are also identified by indicators before income (specifically indicators hypothesized to measure cognitive difficulty). It is clear that income nonresponse and bracketed response can be predicted before they occur. The specificity of these predictions will be the goal future research.

Chapter 5

Summary of Findings and Implications for Research on Income Data Quality and Respondent Verbal Paradata

5.1 Review of Findings

This dissertation presented data that support the general hypothesis that question sensitivity and complexity influence respondent verbal paradata, and that income nonresponse is related to verbal paradata before and at the income question. The effects of question characteristics on verbal paradata will be reviewed, followed by a summary of results showing relationships between verbal paradata with income item nonresponse. Applications of the findings to survey practice, limitations of the research design and future research avenues will also be addressed.

5.1.1 Effects of Question Characteristics

The predicted effects of question characteristics on verbal paradata were largely not supported. Rather than the hypothesized affect/difficulty heightening mechanism that was proposed, sensitivity and complexity seem to lead to less verbal paradata. The true level of physiological and psychological activation that result from question sensitivity and complexity cannot be known in this study. However, it can be assumed that affect and difficulty were stimulated as predicted (sensitive questions lead to heightened subjective experiences of affect and complex questions lead to heightened subjective experiences of difficulty). With this assumption, the results show that these heightened subjective states tend to lead to reduced verbal paradata. Reduced verbal paradata due to heighted psychological states may be caused by 1) extra attention and psychological energy spent on question demands, thus reducing the amount available for verbalization of any kind, or 2) consciously talking less in an effort to get past hard or threatening question as quickly as possible. The cognitive resource explanation is more parsimonious than the explanation citing explicit reduction conversationality, but neither can be tested with these data.

Yet another alternate interpretation of the findings re-characterizes the indicators of affect and difficulty as indicators of "conversationality" more generally (e.g., the degree to which a respondent wants to talk to the interviewer, irrespective of the reason). Under this interpretation, higher rates of indicators (e.g., increased conversationality) on undemanding questions makes intuitive sense. When questions are easy, respondents are willing to talk more. When they are either hard or sensitive (or both), they want to talk less.

Sensitivity and complexity both affected affect and difficulty indicators, suggesting that either the indicators were not accurately assigned to their constructs, or question sensitivity and complexity can each activate both affect and difficulty differently. Further, the same indicator might indicate affective states when present on a sensitive question, and cognitive difficulty when present on a complex question. For example, reports could indicate difficulty coming up with an answer to a difficult question, or could indicate unwillingness to answer a sensitive or threatening question. This notion could be explored further by examining reports coded in these in more detail.

5.1.2 Paradata and Income Nonresponse

Income nonresponse was also related to respondent paradata. Using individual indicators before the income question, differences between nonrespondent types were identifiable before the income question. More specifically, it seems that respondents who eventually provide incomplete income data are identifiable by their behavior, speech, and voice before the income question, compared to those who provide complete income data. Income nonrespondents are identified primarily by affective indicators, suggesting income sensitivity may be causing income nonresponse. Bracketed respondents are identified by cognitive difficulty, which reflects and extends the idea that those who use brackets do so because they find reporting income difficult.

When data at the income question are included as well, income nonresponse can be predicted by a number of indicators, including less backchanneling, more overspeech, lower affect (intensity and valence), more negative comments, more reports and refusals, and less question-answering. Bracketed response is predicted only by more reports, refusals, and more cognitive difficulty. Reports and refusals both identify individuals who do not report full income information, whether income nonresponse or bracketed response, and so their utility in practice is limited. They can predict that the respondent will have some sort of problem reporting income data, but cannot predict whether the respondent will accept income brackets or refuse completely. Other indicators are clearer predictors of income nonresponse, such as affect, backchanneling, and overspeech. Cognitive difficulty seems to be a clear indicator of bracketed response.

Although a measurement model could not be established for the relationship between indicators of affect and difficulty, a one-factor-per-question regression model showed that income nonrespondents are identifiable only by paradata on a nonsensitive, noncomplex question before the income question, while bracketed respondents are identifiable only by paradata at the income question. The placement of these questions may be more interesting than their qualities. Income nonresponse is predicted by factor scores *before* the income question, while bracketed response is predicted by factor scores *at* the income question. This finding has implications for theory of income nonresponse, suggesting perhaps that income nonrespondents have a pre-defined nonresponse propensity, while bracketed respondents make the decision to use brackets in the moment. In terms of interviewing practice, it shows that income nonrespondents can be identified prior to the question, while bracketed respondents are identified more clearly at the question.

Referring to the guiding model (Figure 1), support can be found in this research for question characteristics as causes of verbal paradata, and for respondents as causes of income nonresponse. It is clear that some cognitive and affective components of respondents' response process are evident in verbal paradata. There is strong evidence that, at least for income nonrespondents, it is characteristics of the respondent (e.g., experiences with earlier parts of the survey or a fixed income nonresponse propensity) rather than the income question itself that predict income nonresponse. It is surely the case that other aspects of the "income nonresponse system", such as question or response format, mode, and interviewer characteristics, affect income nonresponse as well. This dissertation only presents evidence for two components of the income nonresponse system.

5.2 Applications of the Results to Survey Practice

The findings presented here should be replicated before application takes place, but are few potential applications are immediate obvious. The first would involve training interviewers to listen for respondent verbal paradata that predict eventual income nonresponse (e.g., negative comments), and training them to intervene to produce better data quality. In this technique, respondent verbal paradata would be treated as diagnostic of later income nonresponse. Interviewers could be trained to do this without modification to a typical survey instrument. The results presented in this dissertation suggest that interviewers should listen for reports, digressions with a codable answer, and increased cognitive difficulty to tell that a respondent might be at a higher propensity for bracketed income response. Bracketed responses could be offered upfront in such cases, or interviewers could simply be prepared to offer them earlier than they would normally. A high number of refusals, and negative comments, as well as more intense and negative affect might suggest that the respondent will refuse to answer income. Interviewers could be prepared to offer confidentiality assurances or statements about the importance of complete data for the research (statements that usually only come after the respondent has refused income). Interviewers could also simply be prepared for resistance from the respondent, and use other ad hoc techniques that might lead to more complete data.

If typical survey questions do not provide enough diagnostic information (e.g., they are not affectively or cognitively demanding enough to produce chances in verbal paradata like those seen here) additional questions could be added to a survey instrument solely for the purpose of providing diagnostic information for prediction of nonresponse. For example, questions that ask respondents to complete complex memory or judgment tasks, ask for exact values of personal financial data or personally identifiable information (e.g., mother's maiden name or social security number) could be added simply to produce more paradata that would, in expectation, identify those respondents at increased risk for providing incomplete income information. Obvious concerns about fatiguing or offending the respondent early in the interview are legitimate, but couching these questions under the guise of "practice questions" might reduce some of these risks.

Interviewers and production-side survey practitioners might argue that the job of the interviewer is difficult enough already, without adding the additional task of listening for verbal paradata and diagnosing nonresponse potential. They are probably right. Some of the verbal paradata analyzed in this dissertation can be processed mechanically (e.g., pitch, pauses, and speech rate in syllables per time unit). As technology to process voice and speech mechanically develops further, reliably identifying fillers, and even specific words may be possible to accomplish in real time. If these paradata can be identified in real time, it seems like a relatively small technical step beyond that to feed the information back to interviewers in the form of an observation and/or instruction (e.g., "This respondent has been highly disfluent. Be sure to offer income brackets"). Such interventions would maximize the use of paradata present in the interview while minimizing additional burden on the interviewer.

From a questionnaire design perspective, the knowledge that question sensitivity and complexity can affect respondent verbal paradata offers another tool for assessing the difficulty and sensitivity of items when designing surveys. While cognitive interviews rely on respondents to tell interviewers that questions are hard or sensitive, verbal paradata from cognitive interviews could be included as part of the data used to assess question problems. For example, it was shown that respondents were less disfluent, among other things, on items that were harder and more sensitive. When pretesting questionnaires, low levels of disfluency and other conversational behavior on items being tested could suggest that they are hard for respondents to answer, and should be changed.

5.3 Limitations and Difficulties with Interpretation of Effects

There are several limitations of the research presented in this dissertation. Some of them have to do with what the study does not address, while others have to do with interpretations of the findings. The study was designed to look at income nonresponse, which is just one component of income data quality. Similarly, it evaluates factors that predict income nonresponse in individual respondents, but does not study nonresponse rates of surveys, or the effect of income nonresponse on total data quality. Applying the current technique and findings to income accuracy would be a worthwhile contribution to the research literature.

With respect to the limitations on inference inherent in the design, first there are limitations in generalizing beyond the survey questions used. Question characteristics (sensitivity and complexity) were only defined by two questions each, leaving room for individual question characteristics to have an unmeasurable effect on the resulting data. Although these questions were selected for their sensitivity and complexity, it is not clear whether they accurately represent other sensitive and complex questions. The range of questions from which to choose was limited to those that were asked in the core section of the Surveys of Consumers. These questions were not designed to be sensitive or complex stimuli. In fact, they did not vary much on these dimensions. Instead of referring to questions as “sensitive” and “nonsensitive”, it may be more accurate to call them "more sensitive" and "less sensitive". Questions tended to be rated as sensitive if they referred to personal income or finances. These topics, while sensitive relative to completely neutral survey questions, are likely not as sensitive as questions about sexual behavior, drug use, or even controversial opinions and attitudes. Other than the income question, none of the questions asked for specific financial values, which likely also reduces their absolute sensitivity. Complex questions seemed to be rated as such based on their request for mathematical calculation or estimation (e.g., "Do you think your income will go up or down more than prices will go up or down…"). Such questions do not tap all types of cognitive complexity (e.g., they do not directly assess knowledge), and so other types complex questions may have different effects on verbal paradata that those found here.

Beyond the questions themselves, the question presentation order was fixed (i.e., Question 1 was the first question for all respondents), so effects of question characteristics cannot be untangled from effects of order if there are any. Similarly, the questions were not evenly spaced. Question 3 came right after Question 2, and both were sensitive questions. The respondent's psychological state at each of these questions is likely to be similar simply because they are close in time. This could explain the obtained effect of sensitivity, rather than any true impact of question sensitivity. These were the limitations of using a pre-existing telephone survey. What was seen as a gain in face validity (i.e., these are real respondents, answering real questions rather than laboratory subjects) was also a sacrifice of scientific purity, and thus the ability to infer causally from the findings. While inferring to a broader population of sensitive and complex questions (and to other types of income questions) is the ultimate goal of research like this, we only really have results describing these five questions "in their natural habitat." The questions came in the same order and roughly at the same time for all respondents, depending only on how long respondents took to answer each question and the few follow-up questions that intervened. There were no major skip patters between these questions. Yet, context effects and fatigue effects may both be present in these data, and they are unexplorable.

Other limitations have to do with the statistical modeling used. Looking at the effect of individual indicators on income nonresponse, two types of models were used; one in which indicators predicted nonresponse, and another in which mean differences between income nonrespondent types were explored. Each is a different model with different assumptions and limitations on the inference of results. The results were mostly compatible, but also differed (recall refusals), so the question of which model provides more accurate and helpful findings for theory and practice remains open. When multivariate modeling was used to measure the factor structure of individual indicators, no sufficient model was found. As a result, the factor scores used to predict income nonresponse are difficult to interpret. This is an imprecise model, and limits the clarity of inference from it.

A larger set of problems with this study involves the endogeneity of the data, specifically, that the data analyzed as outcomes (e.g., verbal paradata and income nonrespondent type) and their predictors are endogenous. That is, they are part of a larger system that may be influenced by other unanalyzed variables, such is interviewer behavior. It is certainly possible that some or all of these analyses suffer from an omitted variable problem. Further analytic steps to follow the dissertation will evaluate the ability of interviewer data to predict respondent voice. Indeed, initial analyses not reported here found interviewer effects on some of the verbal paradata that respondents produce.

A related endogeneity concern has to do with the direction of causation of verbal paradata and income nonresponse. It was shown in this analysis that paradata predict income nonresponse status. But it does not make logical sense to say that the paradata *caused* the nonresponse. The attempt at building a structural equation model was a move toward understanding factors *represented by* verbal paradata that could be causally attributed to nonresponse. Yet, even the development of a good latent variable model using only respondent data would not answer the question of causality if there is an omitted variable problem. There is a larger endogeneity problem to which there is no immediate answer, "how large and complex are the causal relationships between question characteristics, respondent psychological states, respondent verbal paradata, interviewer psychological states, interviewer verbal paradata, and income nonresponse?" Considering the entire system, there are likely multiple causal effects that contribute to income nonresponse (e.g., Figure 1). Some of these will be localized, dyadic causal relationships (e.g., feedback loops), in which something the respondent says (or how they say it) causes the interviewer to change what they say (or how they say it), which in turns leads the respondent to provide an answer they wouldn't have given otherwise. For example, consider this exchange:

I: What was your income *last*-

R: *$3000*

I: *year?* 2009.

R: Oh! For the whole *year*? That's going to be complicated to answer. I work three jobs and do some freelancing. Let's see, there's-

I: Your best estimate would be fine.

R: In that case, $60000.

The respondent starts to answer with their salary last month. The interviewer could have taken this answer but persists in finishing the question. The respondent hears the entire question, admits misunderstanding it, and expresses confusion about how to answer. The respondent then seems to start answering, listing income sources. The interviewer interrupts with a probe telling the respondent that an exact dollar value is not required. The respondent then gives an answer that seems to be an estimate. Would the respondent have added up a more exact value if they were given the time to? Would a different interviewer have accepted the report of $3000? It can't be known, but it seems evident from this fictitious (though not extraordinary) example, that the respondent and interviewer were both influenced by the behavior of the other (e.g., finishing the question, expressing trouble answering, and the explicit acceptance of a less-than-exact answer). Discrete causal connections like this are buried within these data, and require more sophisticated dynamic modeling to address, such as dyadic modeling or micro-level (e.g.,. within-question) time series analysis. It is the long term goal of this project to explore such intricacies of the data.

5.4 Future Directions and Extensions

The future directions of this line of research fall into two categories. The first includes research that can be done on the data collected for the dissertation. The second includes research that goes beyond the dissertation data set. Research that can be done with the dissertation data set includes analysis of interviewer effects, measurement of within-question variability for respondents and interviewers independently, sequential analysis of respondent and interviewer utterances within questions, dyadic analysis of exchanges between interviewers and respondents, and formulation of a better structural model to describe respondent (and perhaps interviewer) psychological states.

Interviewers are a very important and active part of the income nonresponse system that was not evaluated in this dissertation. Future work with these data will examine the effects of interviewers in several ways. Interviewer effects on respondent paradata and income nonresponse will both be explored. Simple fixed interviewer effects, such as the effect of male and female interviewers, old and younger interviewers, and more or less experienced interviewers will be examined. Random effects of interviewers can be explored as well. Finally, and perhaps the largest advance over contemporary research on interviewer effects, will be exploring the relationship of interviewer verbal paradata and respondent verbal paradata at the utterance level, and the eventual influence of these exchanges on later income item nonresponse. Dynamic and dyadic, yet quantitative analyses like these are rare in social science generally and research on interviewer-respondent interaction specifically. The dissertation data provide a perfect source for developing such models. The sequential structure of utterances within each question can also be analyzed in Sequence Viewer by calculating probabilities that certain behaviors occur in conjunction, either immediately together or separated by some number of utterances. For example, it would be possible to calculate the proportion of times that a respondent's expression of confusion is followed by an interviewer's neutral probe. Such events could be used to simply describe the interaction, or to predict the propensity for income nonresponse, as was done with individual indicators and broad factors in the current research. Further, better indicators of affect and cognitive difficulty may be gathered from variables like these that describe the sequence of paradata within questions. For example, are explicit refusals more or less predictive of nonresponse when they co-occur with don't know responses? Does the order of actions matter (e.g., "Oh, I don't know. I don't want to answer," versus "I don't want to answer that…I don't know the answer."). Though not fully dynamic, this kind of analysis gets closer to modeling the actual interaction between interviewers and respondents.

It was also seen that determining a factor structure with these data was difficult. Specific ideas for further exploration of the data's factor structure have been proposed during development of the project. One alternative method involves correlating the individual indicators with question sensitivity and complexity, rather than correlating them with each other. The survey questions have already been rated and ranked on these dimensions, and these classifications can serve as a gold standard against which to compare the presence of verbal paradata. Those items that correlate at least moderately (e.g., .3) with question characteristics could be used to build a measurement model to summarize the data. Another option involves more explicit and decisive characterization of indicators as measures of latent variables. More refined a priori assignments of indicators to constructs could be made, taking into account the types of cognitive difficulty and affect being measured. For example, negative comments and laughing are both signs of affect, but they might not be correlated enough to reveal their common factor. That is, they may be measuring different aspects of affect (e.g., frustration and enjoyment) that are not summarized well by a general "affect" factor that includes both indicators. More exploration of this type is worthwhile considering the theoretical payoff of a factor model that describes the data well.

Beyond the data at hand, future data collection should include experiments designed to test the mechanisms relating question characteristics, respondents' psychological states, verbal paradata, and income nonresponse. Questions could be selected or developed that more strongly manipulate sensitivity and cognitive complexity. Survey questions' effects on verbal paradata could be compared with other psychologically demanding tasks, such as doing mathematical calculation or remembering a complex number. Choosing stimuli that are thought to produce paradata via a variety of mechanisms would help determine what specifically about survey questions leads to higher or lower rates of verbal paradata (e.g., the mechanism behind the paradata).

Finally, this dissertation only dealt with income nonresponse. It would be interesting to replicate these findings with data that could provide measures of response accuracy. It's possible that some of the same factors that predict income nonresponse predict income inaccuracy as well. It is also possible that a completely different set or combination of factors predict income inaccuracy.

Beyond future research, there is conceptual and theoretical work to do to further our understanding of the psychology of survey response (and nonresponse) and interviewer-respondent interaction. The results presented in this dissertation help expand the conceptual framework of verbal paradata, and what they mean for models of data quality. Building on Figure 1, more theoretical and empirical work could develop this framework into more complete theory about the effects of questions on respondents, and the role of respondent paradata in predicting income nonresponse and data quality more broadly.

With respect to survey practice, the findings encourage the idea that interviewers could be trained to explicitly notice respondents’ verbal paradata and intervene proactively to reduce income nonresponse. For example, it was found that respondents who eventually do not provide income make more negative comments before the income question. This is a respondent behavior that should be very salient to the interviewer. There may be an intervention that could be administered to negatively commenting respondents (e.g., a light-hearted comment or re-assurance of the importance of the research) that would reduce their likelihood of income nonresponse. Future studies should explore experimental interventions based on these observational results.

The results presented here open more new questions about the relationships between respondents’ psychological states, verbal paradata, and survey response than they answer. Yet there is ample evidence presented here that further exploration into the dynamic nature of the quality of data collected by interview will be fruitful.

Appendices

Appendix A: Practice Phase Transcription Protocol

Jans Dissertation Transcription Protocol (Update 9-24-08):

Use the following protocol when transcribing audio to text. The goal here is to capture the words that are spoken as well as a few paralinguistic features of the speech. The transcription system is based on Schober & Conrad (1997) and Schober, Conrad, & Fricker (2004).

Text: Transcribe words exactly as spoken. Do not modify words for odd pronunciations or phonetic variation, just use common dictionary spellings.

Marking Turns: Conversation is naturally demarcated by "turns" which are defined by which speaker has the floor. A lot of times turns are clear, but other times they are messy (overlapping speech, interruptions, silence, etc). In survey research we expect a "paradigmatic question-answer sequence" in which the interviewer begins to ask a question, the respondent waits for the question to be finished, and then provides an acceptable answer. The interviewer acknowledges receipt of the answer and/or reads the next question. As simple to accomplish as this may seem, it doesn't always happen. We want our transcription to note where and how the conversational turns took place.

Paralinguistic Markup: Use the following codes for marking-up the text that you've transcribed. See the examples below about how they would be used.

Overlapping speech: Enclose in asterisks and place speech of each partner on a different line.

Example: R interrupts interviewer and the words “survey” and “What are” overlap.

I: Let me start the *survey*.

R: *What are* these questions about?

Example 2: R interrupts I but I finishes their sentence.

I: The survey is *conducted by*-

R: *who*

I: the University of Michigan.

Example 3: R and I talk at the same time, but neither interrupts the other.

I: *I think we*

R: *Well let’s*

Example 4: A more complex example of overlapping speech

R: Um . well she gets her tuitions reduced every year, it looks like *so*

I: *hmm* Well *I don’t-*

R: *And she* works every week for ten hours.

Pauses: Pauses of 1 second or more in length should be marked with a period between to spaces (.). Only pauses that you think are 1-second in length or more need to be marked in transcription. Don’t time it, just estimate. We’ll also get exact measures of pauses in Sequence Viewer.

Interruptions and Restarts: A hyphen should be used where a speaker is cut off mid-word or mid-sentence. Overlapping speech and interruptions may often occur together, but they won’t always. Also, a speaker can interrupt himself or herself (see example 2)

Example 1: R interrupts I but I finishes their sentence.

I: The survey is *conducted by-*

R: *who*

I: the University of Michigan.

Example 2: Interviewer starts and restarts the question

I: When did you la- . sorry . When did he last see a doctor?

Lengthened Sound: A colon (:) in the middle of a word indicates lengthened sound. This is equivalent to repeating a vowel as you might be used to doing to suggest lengthened sound (as in "it's been a looooooong time").

Example:

I: When was the last time you saw a doctor?

R: Oh I guess it's been . u:m . thr:ee years?

Rising Intonation: Rising intonation at the end of an utterance is noted by a "?", and flat or falling intonation is noted by a "."; Use these punctuation marks as indicators of intonation, regardless of whether the utterance was a question or statement. You will likely have many true questions that don't have a question mark at the end. You will also have declarative statements that do.

Appendix B: Practice Phase Coding Scheme

Column	Code Variable	Event Code
1	Actor	I=Interviewer R=Respondent O=Other Person
2	Actor's Behavior	1=Interviewer reads question 2=Interviewer re-reads question 3=Interviewer probes for answer 4=Interviewer re-directs R to task 5=Interviewer reads followup question 6=Respondent answers question 7=Respondent refuses to answer 8=Digression 0=Codable answer with Digression 9=Uncodable behavior
3	Respondent comment (comment on the interview, interaction, questions, task, etc)	0=Not present 1=Negative comment 2=Neutral comment 3=Positive comment 9=Interviewer Turn
4	Pause	0=Not present (no pause) 1=Mid-utterance, empty 2=Mid-utterance, filled 3=Between-utterance, empty 4=Between-utterance, filled 5= Mid-utterance empty and Between-utterance empty 6= Mid-utterance filled and between-utterance empty 7= Mid-utterance empty and between-utterance filled 8= Mid-utterance filled and between-utterance filled
5	Report (Respondent Only)*	0=Not present 1=Present 9=Interviewer Turn
6	Repair	0=Not present 1=Present

7	Anxiety	0-4 (no anxiety to high anxiety)
8	Cognitive Difficulty (Respondent Only)	0-4 (no difficult to high difficulty) 9=Interviewer Turn
9	Interviewer Professionalism	0-4 (low professionalism to high professionalism) 9=Respondent Turn

***All codes are applicable to both Interviewer AND Respondent unless otherwise noted**

<u>Details of How to Use Codes in Table 1</u>

Column 1: Actor

In this column, simply code who is speaking.

Column 2: Actor's Behavior

In this column you'll code both interviewer and respondent behavior using the appropriate code.

1=Interviewer reads question: Reading the question will always be the first interviewer turn.

2=Interviewer re-reads question or response options: If the interviewer reads the question or response options again, code as a re-read

3=Interviewer probes for answer: Probes include things like

> "Whatever it means to you"
> "There's no right or wrong answer"
> "What do you think"
> "What's your best estimate"

When the interviewer asks "Why do you say that", or "Any other reasons" this is really a follow-up question to the first question in the sequence. Use code 5 for this.

4=Interviewer re-directs R to task: Some things interviewers say to get respondents back on-track won't fall into the above categories (i.e., re-read or probe). Code any behavior of the interviewer that you can't code otherwise as an attempt to get the respondent back on track.

5=Interviewer reads follow-up question: This code is used when the interviewer reads the follow-up question built into some of our sequences, such as "Why do you say that" or "Can you tell me more about that" or "Are there any other reasons?"

6=Respondent answers question: If the respondent simply answers the question, use this code

7=Respondent refuses to answer: If the respondent says they don't know or can't answer or anything like that, use this code. This need not be a final response. The interviewer may follow-up with a probe, and then the respondent gives an answer. This code should be used for initial refusals/don't knows and final refusals/don't knows.

8=Digression: Any conversation not directly part of answering the question (on the part of the respondent) gets this code. Any conversation not part of reading questions, recording answers, or getting respondent back from a digression would get this code (i.e., if the interviewer takes part in the digression, rather than getting the respondent back on track).

0=Codable response with digression: Use this code if the respondent answers the question appropriately but also digresses.

9=Uncodable behavior: If the actor's behavior doesn't fit any of these categories, use 9.

Column 3: Respondent comment (comment on the interview, interaction, questions, task, etc)

In this column, code the content of any comment by the respondent on a negative to positive continuum. This variable will likely often be used in conjunction with "Digression" in Column 2. If no comment is present, code 0. An example of a negative comment is "This question is hard" or "This survey is long and boring". Positive comments would include "I like doing this survey" or "It's fun answering these questions". A neutral comment would be something like "How long is this going to take". Use your judgment when in doubt and record anything you'd like to discuss at our meeting.

Use 9 if the turn is an interviewer turn.

Column 4: Pause

Use this column to code whether a pause occurs anywhere in the turn. Remember that a turn starts when the next speaker speaks. So a pause between turns will be associated with the current speaker, not the next speaker.

Empty pauses consist of complete silence, typing or breathing. Filled pauses include some sort of “filler”, such as “um”, “hmmm”, “mhmmm”, or “uhhh”.

Utterances are chunks of speech within turns. One turn can have multiple utterances. You can think of utterances as sentences. If a pause is within an utterance give it a 1 or a 2 depending on whether it is filled. Codes 3 or 4 are pauses between utterances.

Use codes 5-8 if multiple pause types are present in the turn.

Use 0 if no pause is present.

Column 5: Reports

In our context, reporting only makes sense as a respondent behavior. Reporting happens when the respondent does not answer the question directly, but rather than says something about their situation, with respect to the question. For example, if the interviewer asks how much the respondent made in the past year, and the respondent says “I make $10 an hour”, this would be a report. A report could also be more vague, such as “I’m comfortable” or “I make enough”. It has to be relevant to the topic of the question.

Column 6: Repairs

You’ve already been transcribing repairs by indicating them with hyphens (though you’ve been using hyphens for interruptions, too). Repairs are when a person starts an utterance, but then stops, and changes something they’ve said. For example, “The data is rep- the data are representative of the US population” is an example of a repair.

Stammers, where the person restarts an utterance, but does not change anything, or interruptions of one speaker by another are not repairs.

Column 7: Anxiety

Here, use a scale of 0-4 where zero is no anxiety and 4 is high anxiety to code the amount of anxiety you hear in the speaker’s voice. We are using a “mood” definition of anxiety, not a clinical definition, so you should not be thinking of anxiety disorders, but rather anxiety, nervousness, or unease that might occur in everyday conversation.

Column 8: Cognitive Difficulty

In this column (on respondent turns only) code the degree of cognitive difficulty you hear in the respondent’s speech and voice. Respondents might have difficulty coming up with an answer whether they answer the question or not. Difficulty could be for any number of reasons (understanding the question, retrieving an answer, etc). Difficulty may or may not be reflected in spoken words (e.g., “I’m having trouble with this” or “I don’t know how to answer”).

Column 9: Interviewer Professionalism:

What is important here is your perception of their professionalism, rather than whether they hold to a set of defined rules. You’ve seen a little about what interviewers are supposed to do (in the training video we watched), but you also probably have a sense of whether the person sounds professional and competent. On our 0-4 scale 0 represents no professionalism (e.g., the worst interviewer you could imagine) and 4 represents high professionalism (the best interviewer you can imagine).

Appendix C: Final Full-Sample Coding Scheme

Column	Code Variable	Event Code
1	Actor	I = Interviewer R = Respondent O = Other Person u = Uncertain
2	Interviewer Behavior (Iwr Behavior)	0 = Respondent Utterance 1 = Reads question or response options 2 = Re-reads question or response options (partial or full) 3 = Non-directive probe for answer 4 = Other clarification (not neutral probe or re-read) 5 = Re-directs R to task 6 = Reads, re-reads, or probes follow-up question 7 = Neutral feedback and positive reinforcement 8 = Active listening “mhmm”, “I see” etc. 9 = Interviewer comments on the interview a = Agreement with something R says (not clarification) b = Disagreement with something R says (not clarification) c = Conversation management otherwise not codable d = Digression p = Iwr proposes answer r = Repeats or paraphrases what respondent said t = Thank you u = Uncertain/Uncodable behavior
3	Respondent Behavior (R Behavior)	0 = Interviewer Utterance 1 = Answers primary question, no qualification 2 = Answers primary question with qualification 3 = Answers follow-up question, no qualification 4 = Answers follow-up question with qualification 5 = Asks for clarification or repeat of question 6 = Expresses uncertainty about question 7 = Expresses uncertainty about answer or difficulty answering 8 = “Don’t know” answer (explicit) 9 = Refuses to answer (explicit) a = Negative comment b = Digression c = Digression with acceptable answer d = Active listening “mhmm”, “I see” etc. e = Conversation management otherwise not codable f = Agrees with something Iwr says g = Disagrees with something Iwr says h = Don’t know implied

		i = Refusal implied u = Uncertain/Uncodable behavior
4	Laughter	1 = Laughter Present 3 = No Laughter Present u = Uncertain
5	Report (Respondent Only)	0 = Interviewer Utterance 1 = Report Present 3 = No Report Present u = Uncertain
6	Repair	1= Repair Only Present 3 = Stammer/Stutter Only Present 5 = Both Present 9 = Neither Present u = Uncertain
7	Affect Intensity	0-9 (no intensity to high intensity) u = Uncertain
8	Affect Valence	0 = Neutral valence n = Negative valence p = Positive valence u = Uncertain
9	Cognitive Difficulty	0 = No difficult 3 = Some difficulty 5 = High difficulty u = Uncertain

<u>Column 1: Actor</u>

In this column, simply code who is speaking.

I = Interviewer
R = Respondent
O = Other Person (e.g., spouse)

Code "u" if you're not certain who is speaking.

<u>Column 2: Interviewer Behavior</u>

In this column you'll code interviewer behavior using the appropriate code below.

0 = Respondent utterance: Use this code when the utterance is a respondent utterance

1 = Reads primary question or response options (partial or full): Use when the interviewer reads the main question or response options for the first time. Use this code even if the question and response options aren't read in full or are interrupted by the respondent. Reading the question will always be the first interviewer utterance. If the interviewer reads the question or response options again, code as a re-read (2)

Note: The following Primary questions are listed.

> **Primary:** "What about the outlook for prices over the next 5 to 10 years? Do you think prices will be higher, about the same, or lower, 5 to 10 years from now?"
>
> **Primary:** "During the next year or two, do you expect that your (family) income will go up more than prices will go up, about the same, or less than prices will go up?"
>
> **Primary:** "During the next 12 months, do you expect your (family) income to be higher or lower than during the past year?"
>
> **Primary:** "Speaking now of the automobile market - do you think the next 12 months or so will be a good time or a bad time to buy a vehicle, such as a car, pickup, van, or sport utility vehicle?"
>
> **Primary:** "In order to get a picture of people's financial situation we need to know the general range of income of all people we interview. Now, thinking about (your/your family's) total income from all sources (including your job), how much did (you/your family) receive in 200X?"

2 = Re-reads question or response options (partial or full): When the interviewer re-reads the question or response options use this code. The re-read may be requested by the respondent or initiated by the interviewer without a request. Use this code even if the question and response options aren't read in full or are interrupted by the respondent.

3 = Nondirective probe for answer: Nondirective or neutral probes for an answer get this code. These include things like:

> "Whatever it means to you."
> "There's no right or wrong answer."
> "What do you think?"
> "What's your best estimate?"
> "We all hope, but what do you think?"

NOTE: When the interviewer asks "Why do you say that", or "Any other reasons" this is really a follow-up question to the first question in the sequence, NOT a probe. Use code 5 for this.

4 = Other clarification (not neutral probe or re-read): If the interviewer offers any kind of feedback or clarification <u>other than</u> a neutral/nondirective probe or re-read, code it here.

This could include offering a definition of a term or interpreting the goal of the question for the R.

This could also include simply confirming a question the respondent has about the survey question.

> "We mean the prices of all goods and services people buy"
> "I think the study directors want to know"
>
> R: "Do they want me to include grocery stores?"
> I: "Yes"

5 = Re-directs R to task: Use this code for any other action R takes to re-direct R to answering the question other than things that are codable in another category.

6 = Reads or re-reads follow-up question(s): Use this for reading or re-reading the questions that follow the main question (such as "Why do you say that?" or "Are there any other reasons?")

Note: The following are follow-up questions.

> **Follow-up:** "Do you mean that prices will go up at the same rate as now, or that prices in general will not go up during the next 5 to 10 years?"
>
> **Follow-up:** "By about what percent per year do you expect prices to go (up/down) on the average, during the next 5 to 10 years?; How many cents on the

dollar per year do you expect prices to go (up/down) on the average, during the next 5 to 10 years?"

Follow-up: "By about what percent do you expect your (family) income to (increase/decrease) during the next 12 months?"

Follow-up: "Why do you say so?

Follow-up: "Are there any other reasons?"

Follow-up: "Income may predict how people feel about the economy, which is why we ask the income of everyone we interview. We have some range categories if you'd prefer."

Follow-up: "Did (you/your family) receive fifty-thousand dollars or more in 200X?"; "Was it above X dollars?"

7 = Neutral feedback and positive reinforcement: Neutral feedback and positive reinforcement are statements like:

"This is helpful for our research"
"It's helpful to know what people think about this"
"It's good to get your opinion on that"

8 =Active Listening (backchannels): As in normal conversation, interviewers often show that they are listening to respondents by making short utterances like "I see", "mhmm" and, "ah" as the respondent is talking. This may look like filled pauses in the transcript, but they are not. Code them as "active listening".

9 = Interviewer comments on the interview:

"We're about half way done"
"I can call back at another time if it's better"
"I know these questions are hard"

a = Agreement with something R says (not clarification): Include only agreements that don't fit under providing clarification.

b = Disagreement with something R says (not clarification): Include only disagreements that don't fit under providing clarification.

c = Conversation management otherwise not codable: Use this code for verbal behavior that is not codable in another category but is part of conversation management by the interviewer. This includes statements like "I'm just typing this in" or "Let me get this down".

d = Digression: Any conversation not part of reading questions, recording answers, probing, active listening, getting respondent back from a digression, or other conversation management (c) would get this code (i.e., if the interviewer takes part in the digression, rather than getting the respondent back on track).

p = Iwr Proposes Answer: This code applies when the interviewer suggests an answer for the respondent, not simply paraphrases what the respondent says expecting confirmation. The answer proposal may be related to something the respondent said, but is more than just a paraphrase.

R: "I guess go up."
I: "Would that be higher?"

"You said before that you don't have investments"

r = Repeats or paraphrases what respondent said: Use this code when the interviewer repeats or paraphrases the answer that the respondent just gave. This happens most when the interviewer is typing in the respondent's verbatim response to the open-ended questions (e.g., responses to "why do you say so?"). Include verification of respondent's answer here (e.g., "You said better, correct?")

t = Thank you: Use this code if only a "Thank you" is given, with no other feedback or any other behavior in the utterance.

u = Uncertain/Uncodable behavior: Use the code for any behavior you are uncertain about how to code, or is not codable with our current codes. . If you are uncertain, be sure to make a note in your coding notes file. Use this code if multiple behaviors happen in the utterance and you're not sure which to code. Make a note of these in your problem spreadsheet, too.

Column 3: Respondent Behavior

In this column you'll code respondent behavior using the appropriate code below.

0 = Interviewer Utterance: Use this code when it is the interviewer's utterance

1 = Answers primary question with no qualification: Use this code when the respondent answers the primary survey question (the first survey question in the sequence) with no qualification.

2 = Answers primary question with qualification: Use this code when the respondent answers the primary survey question (the first survey question in the sequence) with qualification. Qualifications include things like "Well, I guess it's about ANSWER", or "I don't know, but ANSWER".

3 = Answers follow-up question with no qualification: Use this code when the respondent answers the follow-up survey question in the sequence ("Why do you say that?") with no qualification.

4 = Answers follow-up question with qualification: Use this code when the respondent answers the follow-up survey question in the sequence (e.g., "Why do you say that?") with qualification. Qualifications include things like "Well, I guess it's about ANSWER", or "I don't know, but ANSWER".

Note: The following Primary and Follow-up questions are listed.

> **Primary:** "What about the outlook for prices over the next 5 to 10 years? Do you think prices will be higher, about the same, or lower, 5 to 10 years from now?"
>
> **Follow-up:** "Do you mean that prices will go up at the same rate as now, or that prices in general will not go up during the next 5 to 10 years?"
>
> **Follow-up:** "By about what percent per year do you expect prices to go (up/down) on the average, during the next 5 to 10 years?; How many cents on the dollar per year do you expect prices to go (up/down) on the average, during the next 5 to 10 years?"
>
> **Primary:** "During the next year or two, do you expect that your (family) income will go up more than prices will go up, about the same, or less than prices will go up?"
>
> **Primary:** "During the next 12 months, do you expect your (family) income to be higher or lower than during the past year?"
>
> **Follow-up:** "By about what percent do you expect your (family) income to (increase/decrease) during the next 12 months?"
>
> **Primary:** "Speaking now of the automobile market - do you think the next 12 months or so will be a good time or a bad time to buy a vehicle, such as a car, pickup, van, or sport utility vehicle?"
>
> **Follow-up:** "Why do you say so?
>
> **Follow-up:** "Are there any other reasons?"
>
> **Primary:** "In order to get a picture of people's financial situation we need to know the general range of income of all people we interview. Now, thinking about (your/your family's) total income from all sources (including your job), how much did (you/your family) receive in 200X?"

Follow-up: "Income may predict how people feel about the economy, which is why we ask the income of everyone we interview. We have some range categories if you'd prefer."

Follow-up: "Did (you/your family) receive fifty-thousand dollars or more in 200X?"
"Was it above X dollars?"

5 = Explicitly asks for clarification or repeat of question: Use this code when the respondent asks for any kind of clarification about the question, including the definition of a word in the question (e.g., "What do you mean by X?"), whether to include certain things in their answer (e.g., "Do I include X?")

"What do you mean by prices?"
"Should I include all my income?"
"So this is about the government?"

Also use when respondent asks for the question to be re-read.

6 = Expresses uncertainty about question: Use this if the respondent expresses uncertainty about the question, **but doesn't ask for clarification.** Examples are:

"I don't know what they mean by 'business'"
"This question doesn't make sense to me"
"Didn't you just ask that question?"

"I: Thinking now of the automobile industry…good time to buy a car, van truck or sport utility vehicle?
R: Well, if I think of cars, it's a good time, and if I think of SUVs, it's a bad time."

The distinction between coding a 5 or a 6 is simply whether the respondent explicitly asks for clarification (code a 5) or if they just express uncertainty (code a 6).

7 = Expresses uncertainty about answer or difficulty answering: The difference between coding a 6 and a 7 is whether the respondent is uncertain about the **question** (code 6) or about **how they should answer it given their situation** (code 7). There may be some gray area and overlap here, so just use your best judgment when coding. An example of uncertainty in answering might be:

"It's complicated, I work 3 jobs each with different salaries…"
"Well the gas prices go up and down so much around here…"
"I can't fit my answer into any of those categories"
"Well, my salary income is easy, but my hourly income is difficult to add up"
"I: Thinking now of the automobile industry…good time to buy a car, van truck or sport utility vehicle?

R: Well, I don't follow car prices so it's hard to answer"

8 = "Don't know" answer (explicit): Use this code when the R's entire utterance is a don't know, or their final answer is a don't know. If they say something like "Well, I don't know, BUT I'd say better times", code it as a 3 (answer with qualification).

The respondent should explicitly say that they don't know.

9 = Refuses to answer (explicit): Use this code for any utterance on which the respondent refuses to answer, even if the interviewer is able to convert their refusal to an answer (i.e., this need not be a final refusal). The interviewer may follow-up with a probe, and then the respondent gives an answer on a later utterance.

The respondent should explicitly refuse to answer.

You will also use this code if the refusal is the respondent's final answer. If the respondent says they can't answer, won't answer, or anything like that (other than a "don't know"), use this code. Code "don't know" responses as 8.

a = Negative comment: Use this code when the respondent makes any negative comment about the survey, the question, etc. Examples are:

"This survey is long and boring"
"Whoever wrote this survey should be shot"

Save this category for things that can't be coded as a 6 or 7 (confusion about question or answer). If you can code it as 6 or 7, do that. "a" should be saved for more general comments.

b = Digression <u>without</u> a codable answer: If the respondent gets off-task, code it as a digression. If the respondent simply states confusion about the question or answer, or asks for clarification DO NOT code as digression, but use the appropriate code above. Save this code for cases where the respondent really gets off topic AND does not provide an answer.

c = Digression <u>with</u> a codable answer: See explanation for code "b". Use code "c" when the respondent gets off-task, but also provides a response that the interviewer accepts as their answer.

d = Active listening, "mhmm", "I see", etc. (back channels): As in normal conversation, respondents often show that they are listening to the interviewer by making short utterances like "I see", "mhmm" and, "ah" as the interviewer is talking. This may look like filled pauses in the transcript, but they are not. Code them as "active listening".

e = Conversation management otherwise not codable: Use this code for verbal behavior that is not codable in another category but is part of conversation management by the respondent. This includes statements like “Go ahead” or “I’m ready”.

f = Agrees with something Iwr says: The respondent states agreement (yes, yup, uh huh, etc that are not backchannels) with a statement the interviewer makes, use this code.

For example, if the interviewer proposes an answer (p on Iwr Behavior) or paraphrases what the respondent says (r on Iwr Behavior), and R simply confirms it, use this code.

Include confirmations to probes, offers of bracketed income questions, etc.

g = Disagrees with something Iwr says: The respondent states disagreement (no, nope, nah) with a statement the interviewer makes, use this code.

If the interviewer proposes an answer (p on Iwr behavior) or paraphrases what the respondent says (r on Iwr Behavior), and R simply rejects it, use this code.

Include confirmations to probes, etc. **DO NOT INCLUDE refusals of bracketed income questions. Those should be coded as an explicit refusal.**

h = Don’t Know Implied: Respondents may answer questions by saying, “I can’t answer” or something similar. If the respondent does not explicitly say “I don’t know” or “I won’t tell you”, use this code if you think a “don’t know” is implied.

> “I can’t answer that. My wife does all the bills”

i = *Refusal Implied:* Respondents may answer questions by saying, “I can’t answer” or something similar. If the respondent does not explicitly say “I don’t know” or “I won’t tell you”, use this code if you think a “refusal” is implied.

> “I don’t like to give that out”

u = Uncertain/Uncodable behavior: Use this code if you are uncertain about how to code the behavior or if the behavior is uncodable given our codes. If you are uncertain, be sure to make a note in your coding notes file.

<u>Column 4: Laughter</u>

1 = Laughter: Laughter is present in the utterance. Use the same code for interviewer and respondent utterances. We’ll know from the Actor column whose laughter it is.

3 = No Laughter: Use this code when no laughter is present in the utterance.

u = Uncertain: Uncertain how to code the laughter.

Column 5: Reports (Respondent Only)

In our context, reporting only makes sense as a respondent behavior. Reporting happens when the respondent does not answer the question directly, but says something about their situation with respect to the question. For example, if the interviewer asks how much the respondent made in the past year, and the respondent says "I make $10 an hour", this would be a report. A report on the same survey item could also be something more vague, such as "I'm comfortable" or "I make enough". A report **has to be relevant to the topic of the question.**

0=Interviewer Utterance: Use this when the utterance is an interviewer utterance. Interviewers do not get a report code.

1=Report Present: Use this code on respondent utterances in which you hear a report.

3=Report Not Present: Use this for respondent utterances were you don't hear a report.

u = Uncertain: Use this code if you're not sure whether what you hear is a report.

Column 6: Repairs

Repairs are when a person starts an utterance, but then stops, and changes something they've said. For example, "The data is rep- the data *are* representative of the US population" is an example of a repair.

Stammers or stutters, where the person repeats a syllable or restarts an utterance, ***but does not change anything***, or interruptions of one speaker by another **are not repairs.**

Example:

Stutter/Stammer *Repair*
"**Y-y-yes** I did buy a **fl- some** furniture"

1=Repair Only Present: Use this code on respondent utterances in which you hear a repair.

3 = Stammer/Stutter Only Present: Use this code on respondent utterances in which you hear a stammer or stutter that isn't a repair by the definition above.

5 = Both Present: Use this code when both a repair AND a stammer/stutter are present in the utterance.

9=Neither Present: Use this for respondent utterances were you don't hear a repair or stammer.

u = Uncertain/Uncodable: Use this code if you're not sure whether what you hear is a repair or stutter

<u>Column 7: Affect Intensity</u>

Affect intensity is the amount of affect (feeling, emotion, etc) that you hear in a speaker's voice. To apply this code listen for the amount of affect intensity you hear in the speaker's voice on each utterance. For this code you're ignoring the type of feeling (nervousness, anxiety, frustration, etc) and whether it's positive or negative (frustrations v. satisfaction).

Use the scale below (from 0-9) where 0 is no affect intensity and 9 is the most intense affect you can imagine.

0	1	2	3	4	5	6	7	8	9
No Intensity									High Intensity

There are a number of qualities that indicate increased affect intensity and can be heard in a speaker's voice. This list is not exhaustive or exclusive. The judgment of intensity should be yours. Some aspects of intensity that coders have mentioned are rate of speech, volume, inflection, tone of voice, etc.

Be sure your codes are based on what you hear in the speaker's voice and not what you think the speaker is feeling. The goal is to keep these codes objective (i.e., coding based on what you hear), and to keep inferences (e.g., assumptions about WHY the speaker may be speaking that way to a minimum). Also, make sure you are coding based on the speaker's voice, not whether you like the speaker, whether you would be irritated or happy if you were interviewing or being interviewed by the speaker, or whether you're frustrated or bored with the coding task. Take a deep breath and code away!

<u>Column 8: Affect Valence</u>

Affect valence (positive or negative) is coded for all utterances that receive an intensity code greater than 0. If you code intensity 0, make this code 0 also (neutral).

Once you've decided the intensity in Column 9. Decide on the valence of the affect expressed (if any). Positively valenced affect includes things like happiness, elation, or sounding "upbeat"; good moods. Negatively valenced affect includes things like frustration, anger, or sounding "down"; bad moods.

0 = Neutral valence (apply only when intensity is 0)
n = Negative valence
p = Positive valence

Be sure your codes are based on what you hear in the speaker's voice and not what you think the speaker is feeling. The goal is to keep these codes objective (i.e., coding based on what you hear), and to keep inferences (e.g., assumptions about WHY the speaker may be speaking that way to a minimum). Also, make sure you are coding based on the speaker's voice, not whether you like the speaker, whether you would be irritated or happy if you were interviewing or being interviewed by the speaker, or whether you're frustrated or bored with the coding task. Take a deep breath and code away!

Column 9: Cognitive Difficulty (3 categories, 0-2)

In this column (for respondent AND interviewer utterances) code the degree of cognitive difficulty you hear in the speaker's speech and voice. Respondents might have difficulty coming up with an answer whether they answer the question or not. Difficulty could be for any number of reasons (understanding the question, retrieving an answer, etc). Difficulty may or may not be reflected in spoken words (e.g., "I'm having trouble with this" or "I don't know how to answer"). Interviewers may have difficulty reading questions, probing respondents, or understanding respondents' answers.

NOTE: This is not a 5-point scale. It's a 3-point scale (you can think of it as 0,1,2). The use of 3 and 5 is just to control keying errors.

0 = No Difficulty: The speaker has no problems answering, and you have no reason to believe (either from spoken words or speech and voice qualities) that they had any trouble completing their utterance. Answers in this category will likely be short and fluent, but length and fluency are not requirements.

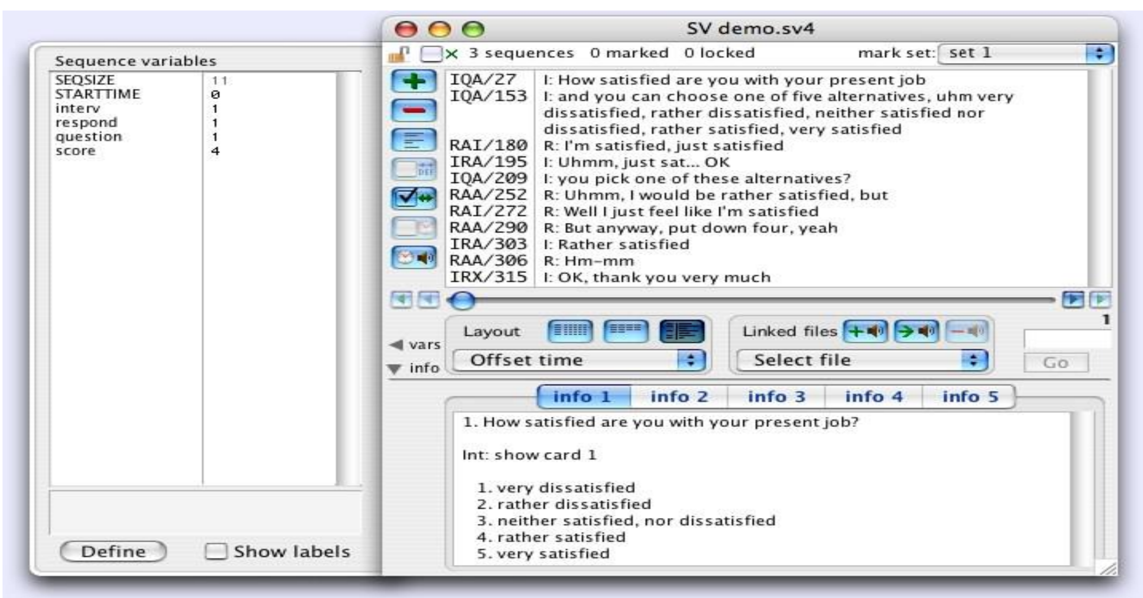
3 = Some Difficulty: You hear a low level of difficulty, but it is not great (i.e., it could be higher). You may notice pausing, fillers, stalling, changes in intonation or speech rate, or explicit statements about having difficulty. These are not requirements for applying a code of 3, nor are they an exhaustive list of things you might hear that suggest difficulty.

5 = Great Difficulty: You hear a high level of difficulty. You may notice more pausing, fillers, stalling, changes in intonation or speech rate, or explicit statements about having difficulty. These are not requirements for applying a code of 2, nor are they an exhaustive list of things you might hear that suggest difficulty.

Appendix D: Overview of Sequence Viewer Program and Use in Coding

The Sequence Viewer software has been developed to code and analyze any phenomena that occur in some sort of temporal sequence, which makes the software particularly useful for studying interpersonal interaction, such as survey interviews. In addition to its analytic functions, Sequence Viewer offers an efficient way to code interpersonal interactions that require the use of a recording, transcript, and numeric coding scheme all at the same time. Below is a screenshot of the Sequence Viewer program. The program displays a transcript text, delimited into "events" (utterances in this case), while the code entry occurs immediately to left of the transcript for easy verification and checking (reducing accidental entry error and facilitating checking and problem resolution relative to unlinked transcripts, codes, and audio files).

Screen Shot of Sequence Viewer Program



An audio file associated with the transcript can be played within the program to aid in coding. Time markings can be added to the transcripts and linked to codes and the associated recordings.

In Sequence Viewer's Terminology, conversational utterances are called "events". In the screen shot above, the events are the lines of the transcript labeled with "I" and "R". Code variables define columns on which each utterance could have one code only. Referring to Appendix C for example, the code variable "Actor" could be coded as "I" for "interviewer, "R" for respondent, or "O" for other speaker. The individual "I" or "R" is referred to as a code. Taking all the columns together (e.g., Actor through Cognitive Difficulty) defines an "event code" in Sequence Viewer terminology. For example, in the screen shot, the event code for the first utterance would be IQA. The digits after the slash are event variables that can also be applied to each event, or calculated from information in the transcript. These types of variables and codes discussed thus far (code variables, codes, event codes, and event variables) describe individual utterances, and each survey question is made up of multiple utterances. In Sequence Viewer, one survey question is displayed per screen, and these screens are referred to as sequences, or sometimes cards. Sequence Variables can be created to describe sequences as well. These can be substantive, but in this project they were mostly used to identify each case (respondent ID and question ID), and coder. In the practice phase, a sequence viewer code for laughter was employed, but that was turned into a code variable (i.e., applied at each utterance) in the full sample coding.

Appendix E: Instructions for Time Stamping Utterances and using Time Keys to Mark Fillers and Pauses

In the next coding step we'll be doing two things that allow us to analyze the role of pauses in survey interviews: assigning time markers (onset and offset times) to spoken utterances and coding filled and pauses in the Time Keys function in SV. Perhaps more than other coding tasks, these require high attention to detail because you will be marking times (i.e., applying codes) on the scale of tenths of seconds.

1) Start by making sure the time units in your SV file are set to 6 ticks (1/10 of a second)
    a. Under File > File Settings > Time check that you're set to 6 ticks, that event time is set to be "relative to start time", and that the box below that is checked.
2) Onset and Offset times are considered "Event Variables" and appear after the slash that follows the series of codes you've been applying. Onset and Offset time will appear in the same space as the sequential utterance counter (if you added that to your file). Only one Event Variable can be visible at a time. You can rotate between event variables.
3) With the Onset time visible, enter 0 for the onset time of the first utterance.
    a. We are defining utterances for the purpose of Onset and Offset times as the duration of spoken words. Onset and Offset times should be set where the person starts speaking in the utterance (Onset) and where they stop speaking (Offset).
    b. If there is a pause between utterances, you will have a gap between Offset time of utterance A and the Onset time of utterance B.
    c. You should never have a situation where the Onset time for B comes before the Offset time for A.
    d. For now, ignore overspeech. Put no time markers on utterances where speakers are talking over each other (i.e., utterances bound by *'s)
    e. Locate the onset time of the second utterance.
        i. Either enter time manually, OR

ii. Click on the play head while it's exactly at the start of the utterance.

f. Continue this through the first 4 non-overlapping utterances.

4) Switch the Event Variable view so that you see "Offset time". Go through steps a-e for offset time of the first 4 non-overlapping utterances. You should then have Onset and Offset time and Offset time for the first 4 non-overlapping utterances.

Shortcuts and Tips:

You can move the play head using the keyboard for more precision than you might get with the mouse.

Onset and offset times
(if 'assign event times' button in main sequence window is checked)
- up arrow: subtracts one unit from event time.
- down arrow: adds one unit to event time.
- Click at event code: assigns present time of sound file to onset or offset time (depending on which one is selected with the popup menu in the layout field).

Control+Up Arrow will play and pause the recording

Using Time Keys to Code Continuous Variables

Time keys are used to code continuous behavior (i.e., behavior that has a presence and absence as well as a duration, such as pausing, gaze (if we were coding video data), body posture, etc).

Under the Keys menu, click "Time Keys"

We'll use the four following two- and three-digit labels to mark the duration of fillers and pauses.

IF (for "Interviewer Filler")

IP (for "Interviewer Pause")

RF (for "Respondent Filler")

RP (for "Respondent Pause")

Once you have used these labels to code Time Keys on one sequence, they may be saved in the SV file, but if not you will have to add them when you set time keys. It's important that they be spelled exactly the same way each time they're used or SV will think that they are different time keys.

Fillers are breaks in speech that contain any kind of "speech-like" token, such as "Um", "Uh", "Hmmm", "Mmmmm", etc (not limited to this list).

Fillers are NOT "active listening/backchannels" as we've been coding already. For example, "Mhm" by the interviewer as a respondent is answering the question is NOT a filler.

Fillers are measures of speech disfluency (like stutters, stammers, repairs, pauses, etc). In the literature and in the project, they are sometimes referred to as "filled pauses", but they don't necessarily need to be "in the middle of a pause".

Wherever the speaker utters a filler (um, uh, ah, etc, NOT active listening) you should enter a time key that covers the duration of that spoken filler only

1) To apply time keys you need to open the Time Keys window from the Keys menu.
2) Make sure your audio file is selected in the "Linked Files" window.
3) When you open the Time Keys window, you see the "view" window.
4) Click on "Edit". If you cannot click "Edit", be sure your audio file is selected.
5) Move the play head (in the Time Keys window) to the point where you think the filler starts.
6) Click on the downward pointing blue arrow next to the question mark to set the beginning of the duration maker.
7) Select the appropriate time key (e.g., IF or RF) or enter a new one in the window that pops up.

8) Move the play head to where you think the pause ends (e.g., where speech begins again).

9) Click on the button with the red circle. A filled-in red line will appear between the start and end points.

10) You can add multiple Time Keys here, but they are not saved until you press "Store".

11) Press "View" to see the color-coded time keys and make sure they look correct.

12) Move on to the next Sequence (next question)

Appendix F: Sequence Viewer Code Variable Reliability Comparisons by Code (category)

F.1 Actor

	Original Coder	Reliability Coder	Equal	Kappa per Category
I=Interviewer	325	325	323	.991
R=Respondent	234	234	232	.990
Total	559	559	555	

F.2 Interviewer Behavior

	Original Coder	Reliability Coder	Equal	Kappa per Category
0 = Respondent Utterance	235	233	230	0.979
1 = Reads question or response options	66	68	65	0.970
2 = Re-reads question or response options (partial or full)	12	16	8	0.571
3 = Non-directive probe for answer	25	23	19	0.791
4 = Other clarification (not neutral probe or re-read)	9	11	5	0.500
5 = Re-directs R to task	2	0	0	0.000
6 = Reads, re-reads, or probes follow-up question	65	68	60	0.901
7 = Neutral feedback and positive reinforcement	11	8	8	0.842
8 = Active listening “mhmm”, “I see” etc.	41	39	38	0.950
9 = Interviewer comments on the interview	0	0	0	0.000
a = Agreement with something R says (not clarification)	1	3	1	0.500
b = Disagreement with something R says (not clarification)	0	0	0	0.000
c = Conversation management otherwise not codable	12	12	8	0.667
d = Digression	0	0	0	0.000
p = Iwr proposes answer	3	2	1	0.400
r = Repeats or paraphrases what respondent said	32	30	29	0.935
t = Thank you	42	39	39	0.963
u = Uncertain/Uncodable behavior	3	7	3	0.600
Total	559	559	514	

F.3 Respondent Behavior

	Original Coder	**Reliability Coder**	**Equal**	**Kappa per Category**
0 = Interviewer Utterance	324	325	321	0.984
1 = Answers primary question, no qualification	36	32	26	0.764
2 = Answers primary question with qualification	32	31	20	0.634
3 = Answers follow-up question, no qualification	43	48	33	0.723
4 = Answers follow-up question with qualification	28	34	22	0.709
5 = Asks for clarification or repeat of question	13	14	10	0.741
6 = Expresses uncertainty about question	3	3	1	0.333
7 = Expresses uncertainty about answer or difficulty answering	3	3	0	0.000
8 = “Don’t know” answer (explicit)	9	10	8	0.842
9 = Refuses to answer (explicit)	6	10	6	0.750
a = Negative comment	1	1	1	1.000
b = Digression	18	14	10	0.625
c = Digression with acceptable answer	11	7	2	0.222
d = Active listening “mhmm”, “I see” etc.	4	2	1	0.333
e = Conversation management otherwise not codable	2	2	2	1.000
f = Agrees with something Iwr says	15	14	11	0.758
g = Disagrees with something Iwr says	4	1	1	0.400
h = Don’t know implied	1	4	0	0.000
i = Refusal implied	2	0	0	0.000
u = Uncertain/Uncodable behavior	4	4	2	0.500
Total	559	559	477	

F.4 Laughter

	Original Coder	**Reliability Coder**	**Equal**	**Kappa per Category**
1 = Laughter Present	12	14	8	0.615
3 = No Laughter Present	547	545	541	0.801
Total	559	559	549	

F.5: Report

	Original Coder	Reliability Coder	Equal	Kappa per Category
0 = Interviewer Utterance	326	328	319	0.963
1 = Report Present	36	25	14	0.457
3 = No Report Present	197	206	177	0.860
Total	559	559	510	

F.6: Repair and Stammer

	Original Coder	Reliability Coder	Equal	Kappa per Category
1= Repair Only Present	13	15	7	.500
3 = Stammer/Stutter Only Present	23	21	15	.681
5 = Both Present	27	10	10	.540
9 = Neither Present	496	513	481	.856
Total	559	559	513	

F.7: Affect Intensity

	Original Coder	Reliability Coder	Equal	Kappa per Category
0 = No Intensity	49	31	1	.021
1	317	562	292	.376
2	206	27	15	.116
3	39	0	0	.000
4	6	1	0	.000
5	4	0	0	.000
6	0	0	0	NA
7	0	0	0	NA
8	0	0	0	NA
9=High Intensity	0	0	0	NA
Total	621	621	294	

F.8: Affect Valence

	Original Coder	Reliability Coder	Equal	Kappa per Category
n=negative	49	31	1	.021
0=neutral	124	160	68	.451
p=positive	448	430	352	.604
Total	621	621	421	

F.9: Cognitive Difficulty

	Original Coder	**Reliability Coder**	**Equal**	**Kappa per Category**
0 = No difficult	491	519	466	.581
3 = Some difficulty	65	37	15	.289
5 = High difficulty	3	3	0	.000
Total	559	559	481	

Appendix G: Univariate Distributions for Indicators by Question

Indicator	**Question 1** **Complex, Nonsensitive**		**Question 2** **Sensitive, Complex**		**Question 3** **Sensitive, Noncomplex**		**Question 4** **Nonsensitive, Noncomplex**		**Question 5** **Income**	
Hypothesized Sensitivity Indicators										
Explicit refusals	Min: .00 Max: .33 Mean: .0085 Med: .0000 Mode: .00 SD:.04743	Skew: 5.699 Kurtosis: 32.201	Min: .00 Max: 0.20 Mean: .0011 Median: .0000 Mode: .00 SD: .01470	Skew: 13.601 Kurtosis: 185.00	Min:.00 Max: .33 Mean: .0018 Median: .0000 Mode: .00 SD: .02451	Skew: 13.601 Kurtosis: 185.00	Min: .00 Max: .14 Mean: .0008 Median: .0000 Mode: .00 SD: .01050	Skew: 13.601 Kurtosis: 185.000	Min: .00 Max: 1.00 Mean: .1654 Median: .0000 Mode: .00 SD: .27676	Skew: 1.822 Kurtosis: 2.476
Implied refusal	Min: .00 Max: .33 Mean: .0080 Median: .0000 Mode: .00 SD: .04731	Skew: 6.228 Kurtosis: 38.776	Min: .00 Max: .00 Mean: .0000 Median: .0000 Mode: .00 SD: .00	Skew: .00 Kurtosis: .00	Min: .00 Max: .67 Mean: .0059 Median: .0000 Mode: .00 SD: .05504	Skew: 10.683 Kurtosis: 120.661	Min: .00 Max: .50 Mean: .0065 Median: .0000 Mode: .00 SD: .05294	Skew: 8.998 Kurtosis: 82.001	Min: .00 Max: 1.00 Mean: .0754 Median: .0000 Mode: .00 SD: .16136	Skew: 2.731 Kurtosis: 8.397
Negative comment about the survey	Min: .00 Max: .20 Mean: .0031 Median: .0000 Mode: .00 SD: .02191	Skew: 7.402 Kurtosis: 56.46	Min:.00 Max: .33 Mean: .0018 Median: .0000 Mode: .00 SD: .02451	Skew: 13.601 Kurtosis: 185.00	Min: .00 Max: .25 Mean: .0027 Median: .0000 Mode: .00 SD: .02298	Skew: 9.149 Kurtosis: 87.992	Min: .00 Max: .11 Mean: .0006 Median: .0000 Mode: .00 SD: .00817	Skew: 13.601 Kurtosis: 185.000	Min: .00 Max: .20 Mean: .0036 Median: .0000 Mode: .00 SD: .02582	Skew: 7.348 Kurtosis: 53.577
Laughs	Min: .00 Max: .40 Mean: .0335 Median: .0000 Mode: .00 SD: .08435	Skew: 2.602 Kurtosis: 6.113	Min: .00 Max: 1.00 Mean: .0368 Median: .0000 Mode: .00 SD: .14153	Skew: 4.731 Kurtosis: 25.155	Min: .00 Max: .67 Mean: .0373 Median: .0000 Mode: .00 SD: .11149	Skew: 3.268 Kurtosis: 10.753	Min: .00 Max: 1.00 Mean: .0688 Median: .0000 Mode: .00 SD: .14671	Skew: 2.935 Kurtosis: 11.143	Min: .00 Max: .67 Mean: .0290 Median: .0000 Mode: .00 SD: .09336	Skew: 3.937 Kurtosis: 17.611
Backchannels	Min: .00 Max: .50 Mean: .0239 Median: .0000 Mode: .00 SD: .07201	Skew: 3.530 Kurtosis: 14.081	Min: .00 Max: .67 Mean: .0152 Median: .0000 Mode: .00 SD: .08228	Skew: 6.416 Kurtosis: 44.558	Min: .00 Max: .25 Mean: .0051 Median: .0000 Mode: .00 SD: .02991	Skew: 6.322 Kurtosis: 41.542	Min: .00 Max: .67 Mean: .0219 Median: .0000 Mode: .00 SD: .07940	Skew: 4.776 Kurtosis: 28.081	Min: .00 Max: .50 Mean: .0345 Median: .0000 Mode: .00 SD: .09449	Skew: 3.260 Kurtosis: 10.956
Conversation management	Min: .00 Max: .27 Mean: .0132 Median: .0000 Mode: .00 SD: .04909	Skew: 3.780 Kurtosis: 13.431	Min: .00 Max: .33 Mean: .0047 Median: .0000 Mode: .00 SD: .03381	Skew: 7.89 Kurtosis: 65.754	Min: .00 Max: .25 Mean: .0101 Median: .0000 Mode: .00 SD: .04122	Skew: 4.368 Kurtosis: 19.090	Min: .00 Max: .33 Mean: .0131 Median: .0000 Mode: .00 SD: .04902	Skew: 4.176 Kurtosis: 18.179	Min: .00 Max: .40 Mean: .0183 Median: .0000 Mode: .00 SD: .06279	Skew: 3.801 Kurtosis: 14.827
Length of question in total utterances	Min: 4 Max: 37 Mean: 11.21 Median: 9.00 Mode: 5 SD: 6.439	Skew: 1.525 Kurtosis: 2.714	Min: 2 Max: 46 Mean: 5.21 Median: 3.00 Mode: 2 SD: 5.509	Skew: 3.820 Kurtosis: 20.323	Min: 2 Max: 25 Mean: 7.36 Median: 6.00 Mode: 5 SD: 4.404	Skew: 1.642 Kurtosis: 2.840	Min: 2 Max: 42 Mean: 12.18 Median: 11.00 Mode: 8 SD: 6.345	Skew: 1.575 Kurtosis: 3.337	Min: 2 Max: 51 Mean: 11.84 Median: 9.00 Mode: 5 SD: 8.589	Skew: 2.120 Kurtosis: 5.726

Indicator	**Question 1 Complex, Nonsensitive**		**Question 2 Sensitive, Complex**		**Question 3 Sensitive, Noncomplex**		**Question 4 Nonsensitive, Noncomplex**		**Question 5 Income**	
Utterances per question (including pauses, Praat)	Min: 5 Max: 11 Mean: 7.63 Median: 8.00 Mode: 8 SD: .888	Skew: -.855 Kurtosis: 2.564	Min: 2 Max: 18 Mean: 5.86 Median: 6.00 Mode: 4 SD: 1.863	Skew: .095 Kurtosis: -1.245	Min: 3 Max: 9 Mean: 7.36 Median: 8.00 Mode: 8 SD: 1.170	Skew: 1.369 Kurtosis: -1.529	Min: 4 Max: 11 Mean: 7.65 Median: 8.00 Mode: 8 SD: .994	Skew: -.971 Kurtosis: 2.943	Min: 4 Max: 12 Mean: 7.85 Median: 8.00 Mode: 8 SD:1.000	Skew: .630 Kurtosis: 5.971
Overspeech	Min: 0 Max: 32 Mean: 4.18 Median: 2.00 Mode: 0 SD: 6.130	Skew: 2.099 Kurtosis: 4.843	Min: 0 Max: 40 Mean: 2.17 Median: .00 Mode: 0 SD: 5.526	Skew: 4.589 Kurtosis: 25.720	Min: 0 Max: 28 Mean: 1.63 Median: .00 Mode: 0 SD: 3.950	Skew: 3.611 Kurtosis: 16.652	Min: 0 Max: 32 Mean: 3.33 Median: .00 Mode: 0 SD: 5.496	Skew: 2.157 Kurtosis: 5.238	Min: 0 Max: 36 Mean: 3.91 Median: .00 Mode: 0 SD: 6.163	Skew: 2.583 Kurtosis: 8.507
Speech rate at first respondent utterance (syllables per second)	Min: .000 Max: 11.765 Mean: 5.21372 Median: 5.0720 Mode: 4.651* SD: 1.905029	Skew: .287 Kurtosis: .674	Min: .000 Max: 11.538 Mean: 4.40745 Median: 4.348 Mode: 2.326* SD: 1.679708	Skew: .412 Kurtosis: 1.145	Min: .612 Max: 12.903 Mean: 4.85089 Median: 4.8245 Mode: 6.061 SD: 2.199025	Skew: .887 Kurtosis: 1.850	Min: .645 Max: 11.364 Mean: 4.34408 Median: 4.34800 Mode: 5.263 SD: 1.698711	Skew: .592 Kurtosis: 1.371	Min: .000 Max: 15.000 Mean: 4.68132 Median: 4.58900 Mode: 3.571 SD: 1.951771	Skew: 1.151 Kurtosis: 4.379
Speechrate (SV)	Min: .16 Max: 14.71 Mean: 1.3965 Median: .7438 Mode: .33* SD: 1.98131	Skew: 3.949 Kurtosis: 19.373	Min: .04 Max: 8.08 Mean: .5341 Median: .3590 Mode: .50 SD: .78001	Skew: 6.373 Kurtosis: 52.034	Min: .06 Max: 20.00 Mean: 1.0043 Median: .4167 Mode: .50 SD: 2.09863	Skew: 5.928 Kurtosis: 43.442	Min: .15 Max: 7.40 Mean: 1.0347 Median: .6165 Mode: 1.40 SD: 1.10854	Skew: 2.713 Kurtosis: 9.222	Min: .09 Max: 27.67 Mean: 1.3274 Median: .6901 Mode: .40 SD: 2.41525	Skew: 7.634 Kurtosis: 77.723
Median f_0 at first respondent utterance	Min: 78.6 Max: 535.4 Mean: 154.969 Median: 150.60 Mode: 86.6 SD: 53.3979	Skew: 2.348 Kurtosis: 13.772	Min: 77.7 Max: 524.2 Mean: 157.741 Median: 159.40 Mode: 92.6 SD: 53.5315	Skew: 2.309 Kurtosis: 12.771	Min: 79.8 Max: 516.3 Mean: 158.751 Median: 157.00 Mode: 93.9 SD: 58.2091	Skew: 2.488 Kurtosis: 11.833	Min: 80.4 Max: 436.6 Mean: 163.619 Median: 160.600 Mode: 103.8 SD: 52.8108	Skew: 1.522 Kurtosis: 5.547	Min: 80.8 Max: 320.9 Mean: 155.839 Median: 154.050 Mode: 98.0* SD: 45.0455	Skew: .537 Kurtosis: .143
5^{th} percentile of f_0 distribution at first respondent utterance	Min: 74.6 Max: 449.3 Mean: 122.780 Median: 117.00 Mode: 75.2 SD:44.1401	Skew: 2.626 Kurtosis: 15.422	Min:74.9 Max: 334.8 Mean: 120.685 Median: 109.60 Mode: 79.7 SD: 38.9656	Skew: 1.464 Kurtosis: 4.191	Min: 74.8 Max: 471.2 Mean: 121.101 Median: 107.60 Mode: 77.7* SD: 43.6841	Skew: 3.124 Kurtosis: 21.319	Min: 75.2 Max: 244.3 Mean: 116.989 Median: 108.500 Mode: 93.1 SD:34.4260	Skew: 1.016 Kurtosis: .723	Min: 75.1 Max: 304.4 Mean: 117.668 Median: 103.750 Mode: 77.0* SD: 38.2519	Skew: 1.325 Kurtosis: 2.545
95^{th} percentile of f_0 distribution at first respondent utterance	Min: 88.5 Max:580.3 Mean: 254.864 Median: 215.80 Mode: 169.7 SD: 129.9842	Skew: 1.100 Kurtosis: .226	Min: 92.3 Max: 589.3 Mean: 262.819 Median: 220.80 Mode: 99.2* SD: 132.6985	Skew: 1.091 Kurtosis: .189	Min: 85.6 Max: 577.9 Mean: 258.604 Median: 218.85 Mode: 165.6* SD:135.3597	Skew: 1.132 Kurtosis: .200	Min: 102.7 Max: 589.1 Mean: 280.756 Median: 248.100 Mode: 248.1* SD: 129.4504	Skew: .879 Kurtosis: -.220	Min: 93.0 Max: 587.6 Mean: 267.830 Median: 225.950 Mode: 106.6* SD: 131.1103	Skew: 1.032 Kurtosis: .021

Indicator	Question 1 Complex, Nonsensitive		Question 2 Sensitive, Complex		Question 3 Sensitive, Noncomplex		Question 4 Nonsensitive, Noncomplex		Question 5 Income	
Difference between 5th and 95th percentile at first respondent utterance	Min: 2.40 Max: 493.10 Mean: 132.08 Median: 75.10 Mode: 19.80* SD: 131.23	Skew: 1.325 Kurtosis: .666	Min: 3.7 Max: 506.50 Mean: 142.13 Median: 96.3 Mode: 18.40* SD: 137.55	Skew: 1.378 Kurtosis: .750	Min: 5.80 Max: 501.70 Mean: 137.50 Median: 80.90 Mode: 42.10 SD:140.07	Skew: 1.307 Kurtosis: .418	Min: 8.0 Max: 513.00 Mean: 163.77 Median: 120.80 Mode: 63.50* SD:137.74	Skew: 1.114 Kurtosis: .179	Min: 2.30 Max: 503.30 Mean: 150.16 Median: 107.1 Mode: 72.00 SD:134.15	Skew: 1.171 Kurtosis: .244
Standard deviation of pitch at first respondent utterance	Min: 1 Max: 208 Mean: 46.09 Median: 30.30 Mode: 6* SD: 45.188	Skew: 1.562 Kurtosis: 2.163	Min: 2 Max: 237 Mean: 51.13 Median: 37.20 Mode: 6* SD: 48.038	Skew: 1.732 Kurtosis: 3.113	Min: 2 Max: 224 Mean: 46.82 Median: 30.40 Mode: 2* SD:45.929	Skew: 1.519 Kurtosis: 1.912	Min: 3 Max: 201 Mean: 57.26 Median: 51.10 Mode: 38 SD: 42.17	Skew: 1.191 Kurtosis: 1.387	Min: 1 Max: 200 Mean: 51.74 Median: 38.85 Mode: 13* SD: 43.380	Skew: 1.321 Kurtosis: 1.496
Pitch (f_0 in Hz) at last 50ms of voicing	Min: 75 Max: 589 Mean: 185.96 Median: 149.80 Mode: 83* SD: 121.872	Skew: 2.057 Kurtosis: 3.462	Min:75 Max: 573 Mean: 185.42 Median: 146.20 Mode: 79* SD: 124.280	Skew: 1.880 Kurtosis: 2.806	Min:76 Max: 593 Mean: 179.33 Median: 151.30 Mode: 78 SD:110.555	Skew: 2.203 Kurtosis: 4.847	Min: 75 Max: 597 Mean: 182.34 Median: 142.20 Mode: 100 SD: 114.314	Skew: 2.101 Kurtosis: 4.178	Min: 75 Max:582 Mean: 183.04 Median: 144.90 Mode: 75 SD: 115.506	Skew: 1.884 Kurtosis: 3.208
Duration of first R utterance (Last f_0 s)	Min: .11 Max: 19.69 Mean: 1.8463 Median: .9250 Mode: .34 SD:2.37785	Skew: 3.429 Kurtosis: 18.028	Min: .03 Max: 21.57 Mean: 2.4312 Median: 1.3950 Mode: .17* SD: 2.83640	Skew: 2.845 Kurtosis: 12.479	Min: .11 Max: 20.73 Mean: 1 Median: 1.8954 Mode: 1.1230 SD: 2.47031	Skew: 3.832 Kurtosis: 21.709	Min: .12 Max: 72.87 Mean: 5.8505 Median: 3.0240 Mode: 2.07* SD: 8.51254	Skew: 4.187 Kurtosis: 25.224	Min: .00 Max: 16.76 Mean: 2.4617 Median: 1.5855 Mode: .71 SD: 2.49634	Skew: 2.707 Kurtosis: 10.360
Duration of first R utterance (R1 Utterance Dur)	Min: .20 Max: 19.80 Mean: 1.9714 Median: 1.0570 Mode: .41* SD: 2.37629	Skew: 3.418 Kurtosis: 18.002	Min: .20 Max: 21.76 Mean: 2.5973 Median: 1.5730 Mode: .43* SD: 2.8472	Skew: 2.854 Kurtosis: 12.438	Min: .24 Max: 20.96 Mean: 2.0210 Median: 1.2395 Mode: .39* SD: 2.47723	Skew: 3.836 Kurtosis: 21.829	Min: .22 Max: 72.99 Mean: 5.9850 Median: 3.1190 Mode: 1.02 SD: 8.51701	Skew: 72.539 Kurtosis: 4.179	Min: .13 Max: 16.85 Mean: 2.6221 Median: 1.7225 Mode: .97* SD: 2.50026	Skew: 2.693 Kurtosis: 10.161
Affect intensity for question	Min: .00 Max: 26.00 Mean: 6.4162 Median: 5.0000 Mode: 4.00 SD: 4.9128	Skew: 1.535 Kurtosis: 2.440	Min: .00 Max: 21.00 Mean: 2.9189 Median: 2.00 Mode: 1.00 SD: 3.47802	Skew: 2.737 Kurtosis: 8.795	Min: .00 Max: 20.00 Mean: 3.9405 Median: 3.0000 Mode: 2.00 SD: 3.34952	Skew: 2.153 Kurtosis: 5.711	Min: .00 Max: 27.00 Mean: 7.1946 Median: 6.0000 Mode: 4.00 SD: 4.89287	Skew: 1.521 Kurtosis: 2.465	Min: .00 Max: 53.00 Mean: 8.7081 Median: 6.0000 Mode: 2.00 SD: 2.00	Skew: 7.97899 Kurtosis: 63.664
Affective valence at utterance (1, 0, -1; positive, neutral, negative)	Min: -1.00 Max: 1.00 Mean: .0393 Median: .0000 Mode: 1.00 SD: .67690	Skew: -.087 Kurtosis: -1.129	Min: -1.00 Max: 1.00 Mean: .1244 Median: .3333 Mode: 1.00 SD: .85924	Skew: -.256 Kurtosis: -1.645	Min:-1.00 Max: 1.00 Mean: .2285 Median: .3333 Mode: 1.00 SD: .80123	Skew: -.365 Kurtosis: -1.447	Min: -1.00 Max: 1.00 Mean: .2408 Median: .4286 Mode: 1.00 SD:.71680	Skew: -.491 Kurtosis: -1.194	Min: -1.00 Max: 1.00 Mean: -.2633 Median: -.5000 Mode: -1.000 SD: .71787	Skew: .515 Kurtosis: .559

Indicator	Question 1 Complex, Nonsensitive		Question 2 Sensitive, Complex		Question 3 Sensitive, Noncomplex		Question 4 Nonsensitive, Noncomplex		Question 5 Income	
Hypothesized Cognitive Complexity Indicators										
Answering primary question	Min: .00 Max: .50 Mean: .1662 Median: .1111 Mode: .00 SD: .18276	Skew: .649 Kurtosis: -.976	Min: .00 Max: 1.00 Mean: .2559 Median: .0000 Mode: .00 SD: .39201	Skew: 1.165 Kurtosis: -.355	Min: .00 Max: 1.00 Mean: .3401 Median: .3333 Mode: .00 SD: .32976	Skew: .755 Kurtosis: -.381	Min: .00 Max: 1.00 Mean: .1498 Median: .0833 Mode: .00 SD: .18978	Skew: 1.593 Kurtosis: 3.577	Min: .00 Max: 1.00 Mean: .0876 Median: .0000 Mode: .00 SD: .17700	Skew: 2.168 Kurtosis: 4.501
Answers primary question with qualification (e.g., I guess, Maybe, etc)	Min: .00 Max: .67 Mean: .1235 Median: .0000 Mode: .00 SD: .16789	Skew:1.185 Kurtosis: .291	Min: .00 Max: 1.00 Mean: .2559 Median: .0000 Mode: .00 SD:.39201	Skew: 1.165 Kurtosis: -.355	Min: .00 Max: 1.00 Mean: .1535 Median: .0000 Mode: .00 SD: .25531	Skew: 1.850 Kurtosis: 3.056	Min: .00 Max: .67 Mean: .1148 Median: .0000 Mode: .00 SD: .15621	Skew: .1.250 Kurtosis: .922	Min: .00 Max: .50 Mean: .0739 Median: .0000 Mode: .00 SD: .15103	Skew: 1.948 Kurtosis: 2.491
Answers follow-up question no qualification	Min: 0 Max: 1.11 Mean: .1970 Median: .1000 Mode: .00 SD: .23309	Skew: 1.004 Kurtosis: .384	Min: .00 Max: .33 Mean: .0026 Median: .0000 Mode: .00 SD: .02661	Skew: 11.355 Kurtosis: 135.311	Min: .00 Max: .67 Mean: .1197 Median: .0000 Mode: .00 SD: .19844	Skew: 1.264 Kurtosis: -.075	Min: .00 Max: .83 Mean: .3551 Median: .3333 Mode: .00 SD: .24837	Skew: .048 Kurtosis: -1.229	Min: .00 Max: .73 Mean: .1095 Median: .0000 Mode: .00 SD: .18209	Skew: 1.532 Kurtosis: 1.286
Answers follow-up question with qualification	Min:.00 Max: .67 Mean: .1825 Median: .1667 Mode: .00 SD: .19343	Skew: .682 Kurtosis: -.654	Min: .00 Max: .43 Mean: .0023 Median: .0000 Mode: .00 SD: .03151	Skew: 13.301 Kurtosis: 185.00	Min:.00 Max: .67 Mean: .1549 Median: .0000 Mode: .00 SD: .20376	Skew: .824 Kurtosis: -.968	Min: .00 Max: .67 Mean: .0462 Median: .0000 Mode: .00 SD: .11698	Skew: 2.854 Kurtosis: 8.292	Min: .00 Max: .50 Mean: .0182 Median: .0000 Mode: .00 SD: .06964	Skew: 4.580 Kurtosis: 22.847
Requests for clarification or repeat of question	Min: .00 Max: .60 Mean: .0389 Median: .0000 Mode: .00 SD: .09785	Skew: 2.737 Kurtosis: 7.891	Min: .00 Max: .67 Mean: .0511 Median: .0000 Mode: .00 SD: .13865	Skew: 2.638 Kurtosis: 5.744	Min:.00 Max: .50 Mean: .0336 Median: .0000 Mode: .00 SD: .09483	Skew: 2.868 Kurtosis: 7.434	Min: .00 Max: .50 Mean: .0256 Median: .0000 Mode: .00 SD: .07420	Skew: 3.355 Kurtosis: 12.679	Min: .00 Max: .60 Mean: .0628 Median: .0000 Mode: .00 SD: .11815	Skew: 1.998 Kurtosis: 3.554
Expression of uncertainty about the question	Min: .00 Max: .33 Mean: .0078 Median: .0000 Mode: .00 SD: .04213	Skew: 5.807 Kurtosis: 34.514	Min: .00 Max: .50 Mean: .0143 Median: .0000 Mode: .00 SD:.07427	Skew: 5.615 Kurtosis: 31.889	Min: .00 Max: .33 Mean: .0064 Median: .0000 Mode: .00 SD: .04126	Skew: 6.925 Kurtosis: 9.118	Min: .00 Max: .60 Mean: .0310 Median: .0000 Mode: .00 SD: .10649	Skew: 3.658 Kurtosis: 12.854	Min: .00 Max: .33 Mean: .0096 Median: .0000 Mode: .00 SD: .04620	Skew: 5.652 Kurtosis: 33.226
Expressions of uncertainty about their answer or how to answer	Min: .00 Max: .40 Mean: .0229 Median: .0000 Mode: .00 SD: .07438	Skew: 3.424 Kurtosis: 11.132	Min: .00 Max: .42 Mean: .0074 Median: .0000 Mode: .00 SD: .04732	Skew: 6.886 Kurtosis: 48.910	Min:.00 Max: .33 Mean: .0069 Median: .0000 Mode: .00 SD: .04106	Min: .00 Max: .60	Min: .00 Max: 1.00 Mean: .0222 Median: .0000 Mode: .00 SD: .09371	Skew: 7.160 Kurtosis: 66.209	Min: .00 Max: .71 Mean: .0201 Median: .0000 Mode: .00 SD: .07612	Skew: 5.807 Kurtosis: 42.389

Indicator	Question 1 Complex, Nonsensitive		Question 2 Sensitive, Complex		Question 3 Sensitive, Noncomplex		Question 4 Nonsensitive, Noncomplex		Question 5 Income	
Explicit "Don't Know" responses	Min: .00 Max: 1.00 Mean: .0591 Median: .0000 Mode: .00 SD: .14487	Skew: 3.216 Kurtosis: 12.574	Min: .00 Max: .55 Mean: .0035 Median: .0000 Mode: .00 SD: .02593	Skew: 8.005 Kurtosis: 66.087	Min: .00 Max: .60 Mean: .0395 Median: .0000 Mode: .00 SD: .11601	Skew: 3.061 Kurtosis: 8.662	Min: .00 Max: 1.00 Mean: .0337 Median: .0000 Mode: .00 SD: .11796	Skew: 4.938 Kurtosis: 30.220	Min: .00 Max: 1.00 Mean: .0646 Median: .0000 Mode: .00 SD: .16407	Skew: 3.269 Kurtosis: 12.401
Implied "Don't Know" responses	Min: .00 Max: .25 Mean: .0097 Median: .0000 Mode: .00 SD: .03964	Skew: 4.433 Kurtosis: 23.221	Min: .00 Max: .00 Mean: .0000 Median: .0000 Mode: .00 SD: .00	Skew: .00 Kurtosis: .00	Min:.00 Max: .33 Mean: .0068 Median: .0000 Mode: .00 SD: .04169	Skew: 6.326 Kurtosis: 40.180	Min: .00 Max: .50 Mean: .0086 Median: .0000 Mode: .00 SD: .05105	Skew: 7.254 Kurtosis: 58.271	Min: .00 Max: .67 Mean: .0272 Median: .0000 Mode: .00 SD: .09105	Skew: 4.049 Kurtosis: 18.959
Digressions	Min: .00 Max: .55 Mean: .0369 Median: .0000 Mode: .00 SD: .09679	Skew: 2.811 Kurtosis: 7.699	Min: .00 Max: 1.00 Mean: .0687 Median: .0000 Mode: .00 SD: .17748	Skew: 2.779 Kurtosis: 7.637	Min: .00 Max: .86 Mean: .0547 Median: .0000 Mode: .00 SD: .14606	Skew: 3.025 Kurtosis: 9.511	Min: .00 Max:1.00 . Mean: .0837 Median: .0000 Mode: .00 SD: .17164	Skew: 2.485 Kurtosis: 6.811	Min: .00 Max: .55 Mean: .0375 Median: .0000 Mode: .00 SD: .10828	Skew: 3.117 Kurtosis: 9.153
Digressions with codable answer	Min: .00 Max: .50 Mean: .0219 Median: .0000 Mode: .00 SD: .07841	Skew: 4.093 Kurtosis: 17.893	Min: .00 Max: 1.00 Mean: .0328 Median: .0000 Mode: .00 SD: .14255	Skew: 4.926 Kurtosis: 26.123	Min: .00 Max: 1.00 Mean: .0261 Median: .0000 Mode: .00 SD: .11188	Skew: 5.522 Kurtosis: 36.502	Min: .00 Max: .50 Mean: .0437 Median: .0000 Mode: .00 SD: .10343	Skew: 2.423 Kurtosis: 5.361	Min: .00 Max: .33 Mean: .0104 Median: .0000 Mode: .00 SD: .04321	Skew: 4.785 Kurtosis: 25.299
Agreement with something Iwr says	Min: .00 Max: .50 Mean: .0416 Median: .0000 Mode: .00 SD: .09456	Skew: 2.482 Kurtosis: 6.353	Min: .00 Max: .50 Mean: .0203 Median: .0000 Mode: .00 SD: .07812	Skew: 4.368 Kurtosis: 19.830	Min: .00 Max: .50 Mean: .0200 Median: .0000 Mode: .00 SD: .07922	Skew: 4.312 Kurtosis: 19.027	Min: .00 Max: .38 Mean: .0304 Median: .0000 Mode: .00 SD: .07344	Skew: 2.571 Kurtosis: 6.342	Min: .00 Max: .75 Mean: .1547 Median: .0000 Mode: .00 SD: .19643	Skew: .923 Kurtosis: -.499
Disagreement with something Iwr says	Min: .00 Max: .20 Mean: .0016 Median: .0000 Mode: .00 SD:.01640	Skew: 10.901 Kurtosis:124.31	Min: .00 Max: .11 Mean: .0006 Median: .0000 Mode: .00 SD: .00816	Skew: 13.601 Kurtosis: 185.00	Min: .00 Max: .10 Mean: .0005 Median: .0000 Mode: .00 SD: .00735	Skew: 13.601 Kurtosis: 185.00	Min: .00 Max: .08 Mean: .0005 Median: .0000 Mode: .00 SD: .00613	Skew: 13.601 Kurtosis: 185.00	Min: .00 Max: .20 Mean: .0069 Median: .0000 Mode: .00 SD: .02944	Skew: 4.543 Kurtosis: 20.914
Reports	Min: .00 Max: .80 Mean: .0744 Median: .0000 Mode: .00 SD: .16295	Skew: 2.321 Kurtosis: 4.833	Min: .00 Max: 1.00 Mean: .1562 Median: .0000 Mode: .00 SD: .29287	Skew: 1.822 Kurtosis: 2.199	Min: .00 Max: 1.00 Mean:.0638 Median: .0000 Mode: .00 SD: .17894	Skew: 3.200 Kurtosis: 10.527	Min: .00 Max: 1.00 Mean: .1653 Median: .0000 Mode: .00 SD: .24514	Skew: 1.482 Kurtosis: 1.529	Min: .00 Max: 1.00 Mean: .0791 Median: .0000 Mode: .00 SD: .17948	Skew: 2.586 Kurtosis: 6.591

Indicator	**Question 1 Complex, Nonsensitive**		**Question 2 Sensitive, Complex**		**Question 3 Sensitive, Noncomplex**		**Question 4 Nonsensitive, Noncomplex**		**Question 5 Income**	
Utterances with a repair only, no stammer	Min: .00 Max: .67 Mean: .0672 Median: .0000 Mode: .00 SD: .13726	Skew: 2.039 Kurtosis: 3.659	Min: .00 Max: 1.00 Mean: .0500 Median: .0000 Mode: .00 SD: .16019	Skew: 4.210 Kurtosis: 19.852	Min: .00 Max: .50 Mean: .0303 Median: .0000 Mode: .00 SD: .09346	Skew: 3.347 Kurtosis: 11.267	Min: .00 Max: 1.00 Mean: .0817 Median: .0000 Mode: .00 SD: .14663	Skew: 2.377 Kurtosis: 8.236	Min: .00 Max: .67 Mean: .0582 Median: .0000 Mode: .00 SD: .13163	Skew: 2.534 Kurtosis: 6.062
Utterances with a stammer only, no repair	Min: .00 Max: .60 Mean: .0531 Median: .0000 Mode: .00 SD: .11631	Skew: 2.488 Kurtosis: 6.221	Min: .00 Max: 1.00 Mean: .0539 Median: .0000 Mode: .00 SD: .16893	Skew: 3.822 Kurtosis: 15.958	Min: .00 Max: 1.00 Mean: .0406 Median: .0000 Mode: .00 SD: .12450	Skew: 4.177 Kurtosis: 22.237	Min: .00 Max: 1.00 Mean: .1065 Median: .0000 Mode: .00 SD: .17518	Skew: 2.250 Kurtosis: 6.886	Min: .00 Max: .50 Mean: .0563 Median: .0000 Mode: .00 SD: .11830	Skew: 2.226 Kurtosis: 4.367
Utterances with a repair and stammer	Min: .00 Max: .50 Mean: .0328 Median: .0000 Mode: .00 SD: .08511	Skew: .2.872 Kurtosis: 8.306	Min: .00 Max: 1.00 Mean: .0270 Median: .0000 Mode: .00 SD: .12770	Skew: 5.906 Kurtosis: 38.462	Min: .00 Max: .50 Mean: .0162 Median: .0000 Mode: .00 SD: .06970	Skew: 4.829 Kurtosis: 24.113	Min: .00 Max: 1.00 Mean: .0844 Median: .0000 Mode: .00 SD: .15850	Skew: 2.800 Kurtosis: 11.133	Min: .00 Max: 1.00 Mean: .0346 Median: .0000 Mode: .00 SD: .11708	Skew: 4.752 Kurtosis: 28.683
Fillers at question	Min: 0 Max: 3 Mean: .52 Median: .00 Mode: 0 SD: .708		Min: 0 Max: 3 Mean: .48 Median: .00 Mode: 0 SD: .692	Skew: 1.331 Kurtosis: 1.161	Min: 0 Max: 4 Mean: .68 Median: .00 Mode: 0 SD: .842	Skew: 1.446 Kurtosis: 2.368	Min: 0 Max: 17 Mean: 1.45 Median: 1.00 Mode: 0 SD: 2.366	Skew: 3.695 Kurtosis: 18.027	Min: 0 Max: 3 Mean: .60 Median: .00 Mode: 0 SD: .768	Skew: 1.190 Kurtosis: .928
Duration of fillers at question	Min: 0 Max: 30 Mean: 3.24 Median: .00 Mode: 0 SD: 5.041	Skew: 2.179 Kurtosis: 6.518	Min: 0 Max: 32 Mean: 2.72 Median: .00 Mode: 0 SD: 4.646	Skew: 2.488 Kurtosis: 9.043	Min: 0 Max: 30 Mean: 4.74 Median: .00 Mode: 0 SD: 6.278	Skew: 1.414 Kurtosis: 1.675	Min: 0 Max: 94 Mean: 8.36 Median: 5.00 Mode: 0 SD: 13.238	Skew: 3.508 Kurtosis: 16.359	Min: 0 Max: 33 Mean: 4.06 Median: .00 Mode: 0 SD: 6.201	Skew: 2.119 Kurtosis: 5.241
Pauses within utterance at question	Min: 0 Max: 7 Mean: .30 Median: .00 Mode: 0 SD:.770	Skew: 4.622 Kurtosis: 31.988	Min: 0 Max: 4 Mean: .24 Median: .00 Mode: 0 SD: .597	Skew: 2.966 Kurtosis: 10.406	Min: 0 Max: 3 Mean: .35 Median: .00 Mode: 0 SD: .650	Skew: 2.146 Kurtosis: 4.873	Min: 0 Max: 20 Mean: 1.44 Median: 1.00 Mode: 0 SD: 2.574	Skew: 4.063 Kurtosis: 22.237	Min: 0 Max: 5 Mean: .30 Median: .00 Mode: 0 SD: .703	Skew: 3.381 Kurtosis: 15.136
Total duration of pauses	Min: 0 Max: 77 Mean: 5..64 Median: .00 Mode: 0 SD: 13.704	Skew: 2.969 Kurtosis: 9.369	Min: 0 Max: 90 Mean: 4.92 Median: .00 Mode: 0 SD: 14.415	Skew: 3.901 Kurtosis: 16.755	Min: 0 Max: 142 Mean: 7.01 Median: .00 Mode: .00 SD: 16.392	Skew: 4.283 Kurtosis: 26.692	Min: 0 Max: 426 Mean: 24.38 Median: 9.00 Mode: 0 SD: 48.069	Skew: 4.859 Kurtosis: 32.140	Min: 0 Max: 92 Mean: 6.19 Median: .00 Mode: 0 SD: 15.427	Skew: 3.263 Kurtosis: 11.777

Indicator	Question 1 Complex, Nonsensitive		Question 2 Sensitive, Complex		Question 3 Sensitive, Noncomplex		Question 4 Nonsensitive, Noncomplex		Question 5 Income	
Total words at question	Min: 45 Max: 390 Mean: 119.96 Median: 100.00 Mode: 61* SD: 60.522	Skew: 1.851 Kurtosis: 3.923	Min: 32 Max: 265 Mean: 63.09 Median: 42.00 Mode: 33 SD: 43.449	Skew: 2.282 Kurtosis: 5.868	Min: 22 Max: 240 Mean: 63.96 Median: 51.00 Mode: 45 SD: 36.860	Skew: 1.879 Kurtosis: 4.136	Min: 52 Max: 479 Mean: 149.56 Median: 121.00 Mode: 75 SD: 85.604	Skew: 1.565 Kurtosis: 2.506	Min: 54 Max: 495 Mean: 129.90 Median: 113.00 Mode: 77 SD: 65.619	Skew: 2.179 Kurtosis: 6.909
Respondent words per utterance	Min: 1.33 Max: 30.80 Mean:6.79 Median: 5.5 Mode: 3.0 SD: 4.83	Skew: 2.088 Kurtosis: 6.020	Min: 1.0 Max: 31.00 Mean: 8.51 Median: 5.33 Mode: 1.0 SD: 5.33	Skew: 1.953 Kurtosis: 5.318	Min: 1.0 Max: 17.25 Mean: 5.74 Median: 5.0 Mode: 2.5 SD: 3.62	Skew: 1.334 Kurtosis: 1.440	Min: 2.0 Max: 85.5 Mean: 14.69 Median: 11.83 Mode: 6.0 SD: 10.56	Skew: 2.462 Kurtosis: 10.467	Min: 1.5 Max: 35.5 Mean: 6.25 Median: 5.0 Mode: 4.0 SD: 4.63	Skew: 3.24 Kurtosis: 14.514
R cognitive difficulty at utterance (0, 1, 2)	Min: .00 Max: 1.50 Mean: .2459 Median: .0000 Mode: .00 SD: .34169	Skew: 1.271 Kurtosis: .791	Min: .00 Max: 2.00 Mean: .2184 Median: .0000 Mode: .00 SD: .38824	Skew: 1.950 Kurtosis: 3.984	Min: .00 Max: 1.00 Mean: .2263 Median: .0000 Mode: .00 SD: .35274	Skew: 1.294 Kurtosis: .179	Min: .00 Max: 1.00 Mean: .2082 Median: .0000 Mode: .00 SD: .28642	Skew: 1.251 Kurtosis: .631	Min: .00 Max: 1.60 Mean: .2395 Median: .0833 Mode: .00 SD: .30895	Skew: 1.369 Kurtosis: 1.814
R had no difficulty at utterance	Min: .00 Max: 1.00 Mean: .7697 Median: 1.0000 Mode: 1.00 SD: .31173	Skew: -1.096 Kurtosis: .012	Min: .00 Max: 1.00 Mean: .7972 Median: 1.0000 Mode: 1.00 SD: .34269	Skew: -1.445 Kurtosis: .609	Min: .00 Max: 1.00 Mean: .7821 Median: 1.0000 Mode: 1.00 SD: .33902	Skew: -1.317 Kurtosis: .327	Min: .00 Max: 1.00 Mean: .7983 Median: 1.0000 Mode: 1.00 SD: .27925	Skew: -1.317 Kurtosis: .934	Min: .00 Max: 1.00 Mean: .7704 Median: .9167 Mode: 1.00 SD: .28790	Skew: -1.087 Kurtosis: .224
R had some difficulty at utterance	Min: .00 Max: 1.00 Mean: .2146 Median: .0000 Mode: .00 SD: .29450	Skew: 1.124 Kurtosis: .121	Min: .00 Max: 1.00 Mean: .1873 Median: .0000 Mode: .00 SD: .32948	Skew: 1.584 Kurtosis: 1.106	Min: .00 Max: 1.00 Mean: .2095 Median: .0000 Mode: .00 SD: .33367	Skew: 1.403 Kurtosis: .600	Min: .00 Max: 1.00 Mean: .1954 Median: .0000 Mode: .00 SD: .27717	Skew: 1.382 Kurtosis: 1.133	Min:.00 Max: 1.00 Mean: .2197 Median: .0769 Mode: .00 SD: .27655	Skew: 1.090 Kurtosis: .263
R had high difficulty at utterance	Min: .00 Max: .50 Mean: .0156 Median: .0000 Mode: .00 SD: .06757	Skew: 5.200 Kurtosis: 30.219	Min: .00 Max: 1.00 Mean: .0155 Median: .0000 Mode: .00 SD: .11050	Skew: .8.137 Kurtosis: 68.907	Min: .00 Max: .50 Mean: .0084 Median: .0000 Mode: .00 SD: .05427	Skew: 7.004 Kurtosis: 51.577	Min: .00 Max: .33 Mean: .0066 Median: .0000 Mode: .00 SD: .03809	Skew: 6.673 Kurtosis: 47.062	Min: .00 Max: .60 Mean: .0099 Median: .0000 Mode: .00 SD: .05548	Skew: 7.850 Kurtosis: 73.476

*Multiple modes exist. Smallest value is shown.

Appendix H: Pearson Correlations of Utterances with Individual Indicators

Bivariate correlations between two measures of interaction lengths (the overall number of utterances and number of respondent utterances) and counts of individual indicators are presented below. The magnitude of the correlations between the count of each indicator and respondent utterances varied widely from r=.035 to .720. Suggesting that some indicators are more likely to be present when exchanges between respondents and interviewers are longer, but some are hardly influenced by interaction length. Relatively low correlations with question length (less than r=.3, e.g.) were found for refusals, answer with qualification, uncertainty about the question, uncertainty about the answer, don't knows, and negative comments. Other than negative comments, these are indicators that we would expect to come early in the interaction, possibly at the first respondent utterance and coincide with short interactions. Refusals, answers, and don't knows are likely to occur only once per question in paradigmatic interactions. The correlation between number of utterances and overspeech was high at r=.72. Each conversational turn provides an opportunity for over speech, the amount of over speech is unlimited. It should be noted, however, that high correlation does not necessarily mean that the indicator occurs later in the question. It simply means that the indicator is more likely to occur on questions that have longer exchanges. The majority of correlations were moderate (r=.3-.5). As might be expected, all indicators were positively correlated with number of utterances (i.e., no indicator was found less often in longer exchanges). The indicator for answers with no qualification was not correlated with question length. Straightforward substantive answer to the question (e.g., not a refusal or don't know, and no qualification to the answer, each of which would give the interviewer reason to probe) would lead to short exchanges (e.g., a restriction in range on the utterance axis) and thus little or no correlation.

The sizable correlations between utterances and indicators suggest that for many of the indicators, a count cannot be used unless the number of utterances is included as a control variable in the model. Acoustic variables are only measured on one respondent utterance, so the number of utterances is not an issue for these. The values of these indicators can be used directly without concern. If indicator presence is truly correlated with question length, it will show up when means/proportions are used, but will not be artificially influenced by count of occurrences. For this reason, all indicators analyzed in the rest of the dissertation are calculated as a mean or proportion at each question, except the "number of utterances" indicator, which can only be a count.

The codes discussed with significant correlations are all defined as presence or absence at each utterance, and so it is intuitive that more utterances would lead to more occurrence. However, the relationship between length and affect valence (coded -1, 0, 1) is too small in magnitude to be meaningful, suggesting that affect is not necessarily related to length. This correlation holds under the coding that was applied at each utterance (three-level, None, Some, High, 0, 1, 2) or whether the binary representations of each category are used.

Utterance duration (the length of each respondent utterance) was also positively, though only slightly correlated with question length (.119 with respondent utterances and .112 with total utterances). Defining question-level indicators as counts of behaviors or phenomena has clear problems.

Correlation of Indicator Counts with Respondent and Total Utterances

	Correlation with…	
Indicator	**Respondent Utterances**	**Total Utterances**
Hypothesized Affect Indicators		
Explicit Refusal	.106 (.001)	.096 (.003)
Implied Refusal	.174 (<.0005)	.155 (<.0005)
Backchannel	.352 (<.0005)	.333 (<.0005)
Conversation Management	.390 (<.0005)	.360 (<.0005)
Laughter	.412 (<.0005)	.380 (<.0005)
Utterance duration	.119 (<.0005)	.112 (.001)
Overspeech	.720 (<.0005)	.734 (<.0005)
Negative comment	.119 (<.0005)	.084 (.011)
Affect intensity	.858 (<.0005)	.823 (<.0005)
Affect valence	.077 (.018)	.099 (.002)
Hypothesized Cognitive Difficulty Indicators		
Answers with qualification	.073 (.027)	.079 (.017)
Answers without qualification	.035 (.288)	.056 (.090)
Request for clarification or repeat	.467 (<.0005)	.435 (<.0005)
Uncertainty about question	.213 (<.0005)	.205 (<.0005)
Uncertainty about answer	.297 (<.0005)	.260 (<.0005)
Explicit don't know	.282 (<.0005)	.228 (<.0005)
Implied don't know	.231 (<.0005)	.198 (<.0005)
Digression with no codable answer	.506 (<.0005)	.498 (<.0005)
Digression with codable answer	.252 (<.0005)	.228 (<.0005)
Report	.340 (<.0005)	.314 (<.0005)
Repair only	.497 (<.0005)	.478 (<.0005)
Stammer only	.422 (<.0005)	.407 (<.0005)
Repair and stammer	.335 (<.0005)	.338 (<.0005)
Cognitive difficulty rating	.426 (<.0005)	.402 (<.0005)
No cognitive difficulty	.879 (<.0005)	.857 (<.0005)
Some cognitive difficulty	.410 (<.0005)	.380 (<.0005)
High cognitive Difficulty	.244 (<.0005)	.247 (<.0005)

Appendix I: Table of F-value for Sensitivity and Complexity Effects for Individual Indicators

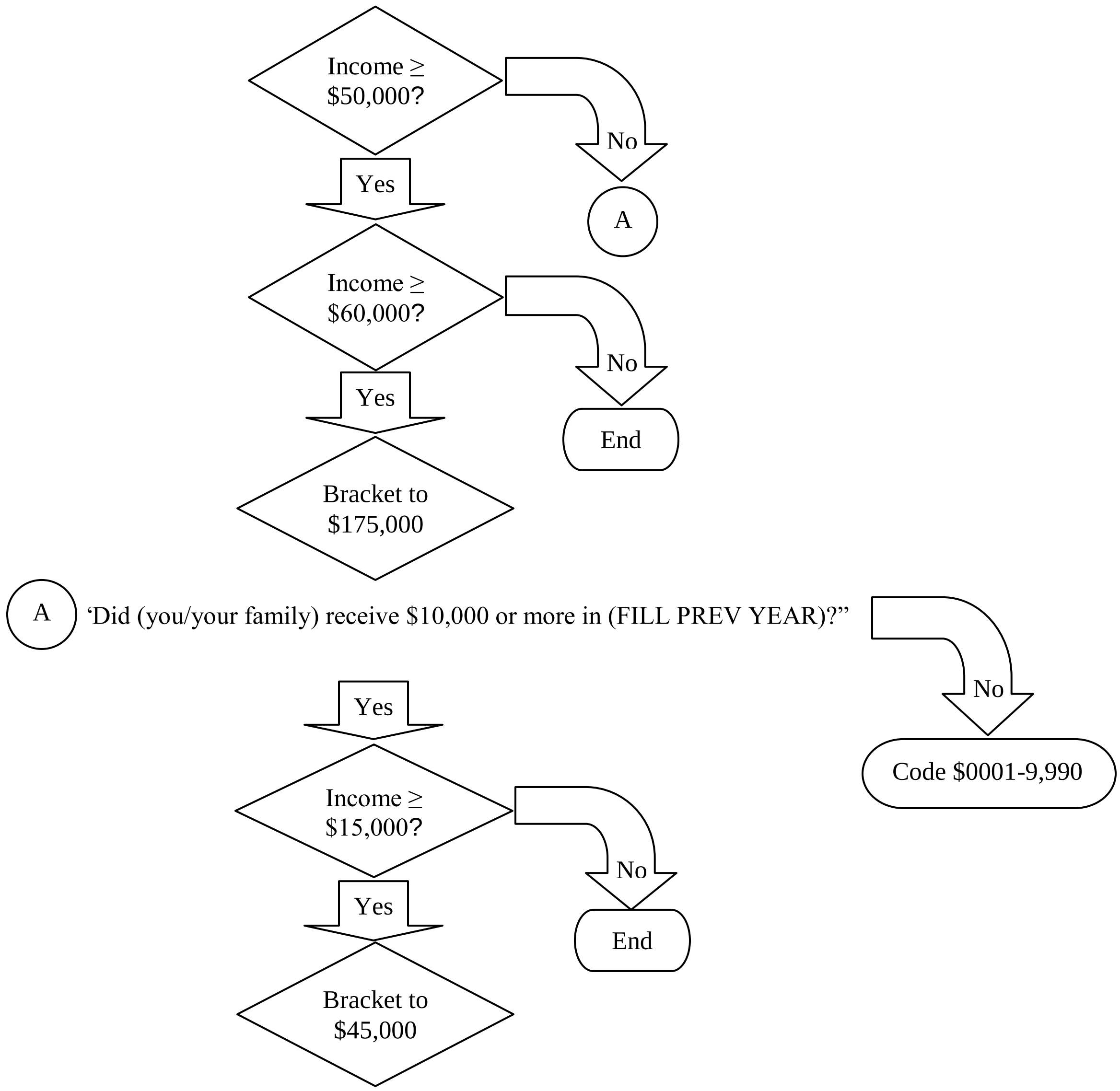
Each cell of the table below presents the F-value (and p-value in parentheses). Degrees of freedom for each test are noted in below the table. The table displays the indicators by their hypothesized relationship to question sensitivity or complexity. Significant main effects and interactions at the α=.05 level are in **bold**.

Indicator	**Sensitivity (S)**	**Complexity (C)**	**S x C**
Hypothesized Sensitivity Indicators			
Explicit refusal	2.37 (.129)	2.809 (.095)	**4.118 (.044)**
Implied refusal	2.399 (.123)	.604 (.438)	.972 (.326)
Having negative comment about the survey	.074 (.786)	.273 (.602)	1.318 (.252)
Laughs	2.918 (.089)	**4.723 (.031)**	**4.033 (.046)**
Backchanneling	**7.436 (.007)**	2.134 (.146)	.632 (.428)
Conversation management	**4.064 (.045)**	.709 (.401)	.909 (.342)
Length of question in total utterances	**218.9 (<.0005)**	**14.912 (<.0005)**	2.449 (.119)
Overspeech per utterance	**7.351 (.007)**	**11.259 (.001)**	.160 (.690)
Speech rate at first respondent utterance (syllables per second)	1.272 (.261)	2.647 (.105)	**24.332 (<.0005)**
Median f_0 at first respondent utterance	.215 (.644)	1.560 (.213)	2.559 (.111)
5^{th} percentile of f_0 distribution at first respondent utterance	.165 (.685)	1.326 (.251)	1.782 (.184)
95^{th} percentile of f_0 distribution at first respondent utterance	.573 (.450)	2.000 (.159)	3.491 (.063)
Difference between 5^{th} and 95^{th} percentile at first respondent utterance	.710 (.401)	3.158 (.077)	**4.632 (.033)**
Standard deviation of pitch at first respondent utterance	.701 (.404)	1.868 (.173)	**6.340 (.013)**
Pitch (f_0 in Hz) at last 50ms of voicing	.010 (.919)	.208 (.649)	.075 (.785)
Duration of first R utterance	**24.407 (<.0005)**	**22.897 (<.0005)**	**50.929 (<.0005)**
Average affect intensity per utterance	**158.7 (<.0005)**	**11.085 (.001)**	.168 (.682)
Average affective valence per utterance (1, 0, -1; positive, neutral, negative)	.627 (.430)	**11.395 (.001)**	1.242 (.267)
Hypothesized Cognitive Complexity Indicators			
Answering primary question	**142.1 (<.0005)**	**20.255 (<.005)**	**11.029 (.001)**
Answering primary question with qualification	**20.887 (<.0005)**	**9.616 (.002)**	**7.038 (.009)**
Request for clarification or repeat of question	2.073 (.152)	**4.311 (.039)**	.069 (.793)

Expression of uncertainty about the question	3.379 (.068)	2.697 (.102)	**7.449 (.007)**
Expresses expressions of uncertainty about their answer or how to answer	**12.362 (.001)**	.015 (.903)	.001 (.977)
Explicit "Don't Know" response	**11.686 (.001)**	.581 (.447)	**12.563 (<.005)**
Implied "Don't Know" response	**4.451 (.036)**	1.036 (.310)	1.769 (.185)
Digresses	.017 (.897)	2.352 (.127)	**8.811 (.003)**
Digressions with codable answer	.197 (.658)	.801 (.372)	3.763 (.054)
Agreement with something the interviewer says	**7.170 (.008)**	.990 (.321)	.982 (.323)
Disagreement with something the interviewer says	.373 (.542)	.653 (.420)	.532 (.467)
Report	.389 (.534)	.003 (.959)	**33.538 (<.0005)**
Repair only, no stammer	**11.560 (.001)**	.073 (.788)	2.856 (.093)
Stammer only, no repair	**8.689 (.004)**	3.285 (.072)	**11.327 (.001)**
Repair and stammer	**19.614 (<.0005)**	**6.164 (.014)**	**16.743 (<.0005)**
Fillers per utterance	**8.684 (.004)**	**27.051 (<.0005)**	**18.319 (<.0005)**
Filler duration per utterance	**4.461 (.036)**	**27.618 (<.0005)**	**9.676 (.002)**
Pauses per utterance	**28.292 (<.0005)**	**33.854 (<.0005)**	**32.601 (<.0005)**
Pause duration per utterance	**16.034 (<.0005)**	**23.582 (<.0005)**	**21.876 (<.0005)**
Total words per utterance	3.137 (.078)	**55.132 (<.0005)**	**199.2 (<.0005)**
Respondent words per utterance	**114.3 (<.0005)**	**61.932 (<.0005)**	**100.5 (<.0005)**
Average rated respondent cognitive difficulty at utterance (0, 1, 2)	.061 (.805)	.689 (.408)	1.413 (.236)
Proportion of utterances with no difficulty	.103 (.749)	.158 (.691)	1.575 (.211)
Proportion of utterances with some difficulty	.135 (.713)	.006 (.940)	1.494 (.223)
Proportion of utterances with high difficulty	.025 (.875)	2.708 (.102)	.035 (.852)

Appendix J: Flow Chart of the Open-ended Income Question and the Series of Brackets Used by the SCA

Open-ended Income Question: "To get a picture of people's financial situation we need to know the general range of income of all people we interview. Now, thinking about (your/your family's) total income from all sources (including your job), how much did (you/your family) receive in (FILL PREVIOUS YEAR)?IF REFUSAL OR DON'T KNOW: "Did (you/your family) receive $50,000 or more in (FILL PREV YEAR)?"



Appendix K: One-way ANOVA Results with Income Nonrespondent Type

Each cell of the table below presents the F-value (and p-value in parentheses). Degrees of freedom for each test are noted in below the table. The table displays the indicators by their hypothesized relationship to question sensitivity or complexity. Significant effects at the α=.05 level are in **bold**.

Indicator*	**Income NR Type Model F-value (p-value)**
Hypothesized Sensitivity Indicators	
Explicit refusal	.410 (.664)
Implied refusal	1.143 (.321)
Negative comment about the survey	**5.6145 (.004)**
Laughs	.736 (.480)
Backchanneling	1.168 (.313)
Conversation management	.030 (.970)
Length of question in total utterances	.004 (.996)
Overspeech per utterance	1.533 (.219)
Speech rate at first respondent utterance (syllables per second)	.475 (.623)
Median f_0 at first respondent utterance	.530 (.589)
5^{th} percentile of f_0 distribution at first respondent utterance	.298 (.742)
95^{th} percentile of f_0 distribution at first respondent utterance	.529 (.590)
Difference between 5^{th} and 95^{th} percentile at first respondent utterance	.099 (.906)
Standard deviation of pitch at first respondent utterance	.106 (.900)
Pitch (f_0 in Hz) at last 50ms of voicing	2.328 (.100)
Duration of first R utterance	.220 (.803)
Average affect intensity per utterance	**6.670 (.002)**
Average affective valence per utterance (1, 0, -1; positive, neutral, negative)	**8.825 (<.0005)**
Hypothesized Cognitive Complexity Indicators	
Answering primary question	.580 (.561)
Answering primary question with qualification	.102 (.903)
Request for clarification or repeat of question	.563 (.571)
Expression of uncertainty about the question	.062 (.940)

Indicator*	**Income NR Type Model F-value (p-value)**
Expresses expressions of uncertainty about their answer or how to answer	*2.594 (.077)*
Explicit “Don’t Know” response	1.028 (.360)
Implied “Don’t Know” response	1.067 (.346)
Digression	**6.042 (.003)**
Digressions with codable answer	**5.128 (.007)**
Agreement with something the interviewer says	*2.514 (.084)*
Disagreement with something the interviewer says	1.312 (.272)
Report	*2.878 (.059)*
Repair only, no stammer	.249 (.780)
Stammer only, no repair	.369 (.692)
Repair and stammer	.393 (.675)
Fillers per utterance	.890 (.413)
Filler duration per utterance	.822 (.441)
Pauses per utterance	.603 (.549)
Pause duration per utterance	.279 (.757)
Total words per utterance	.189 (.828)
Respondent words per utterance	.003 (.997)
Average rated respondent cognitive difficulty at utterance (0, 1, 2)	**32.164 (<.0005)**
No difficulty	**34.102 (<.0005)**
Some difficulty	**34.427 (<.0005)**
High difficulty	**3.058 (.049)**

Income NR Type: $F(2, 182)$

*Unless otherwise stated, each indicator is the average across four items of the proportion of utterances at each item on which the indicator occurred.

Appendix L: Multinomial Logistic Regression of Income Nonrespondent Type on Individual Indicators

The table below includes all the predictors used in the multinomial logistic regression, their coefficients, and the standard error and significance for each coefficient. The first half of the table presents coefficients for prediction of the log odds of income nonresponse relative to dollar amount response (i.e., the regression of income nonresponse on each predictor). The second half of the table presents the coefficients for the regression of bracketed response on each predictor.

The multinomial logistic regression predicts income nonrespondent type, using dollar amount response as the reference category. Coefficients represent the increase in the log odds of being a nonrespondent (or bracketed respondent) relative to a dollar amount respondent for a one unit increase in that predictor, given all other predictors in the model are held constant. All p-values are 2-tailed, and coefficients significant at the α=.05 level are in bold.

Coefficients, Standard Errors, and p-values from Multinomial Logistic Regression of Income Nonrespondent Type on Verbal Paradata

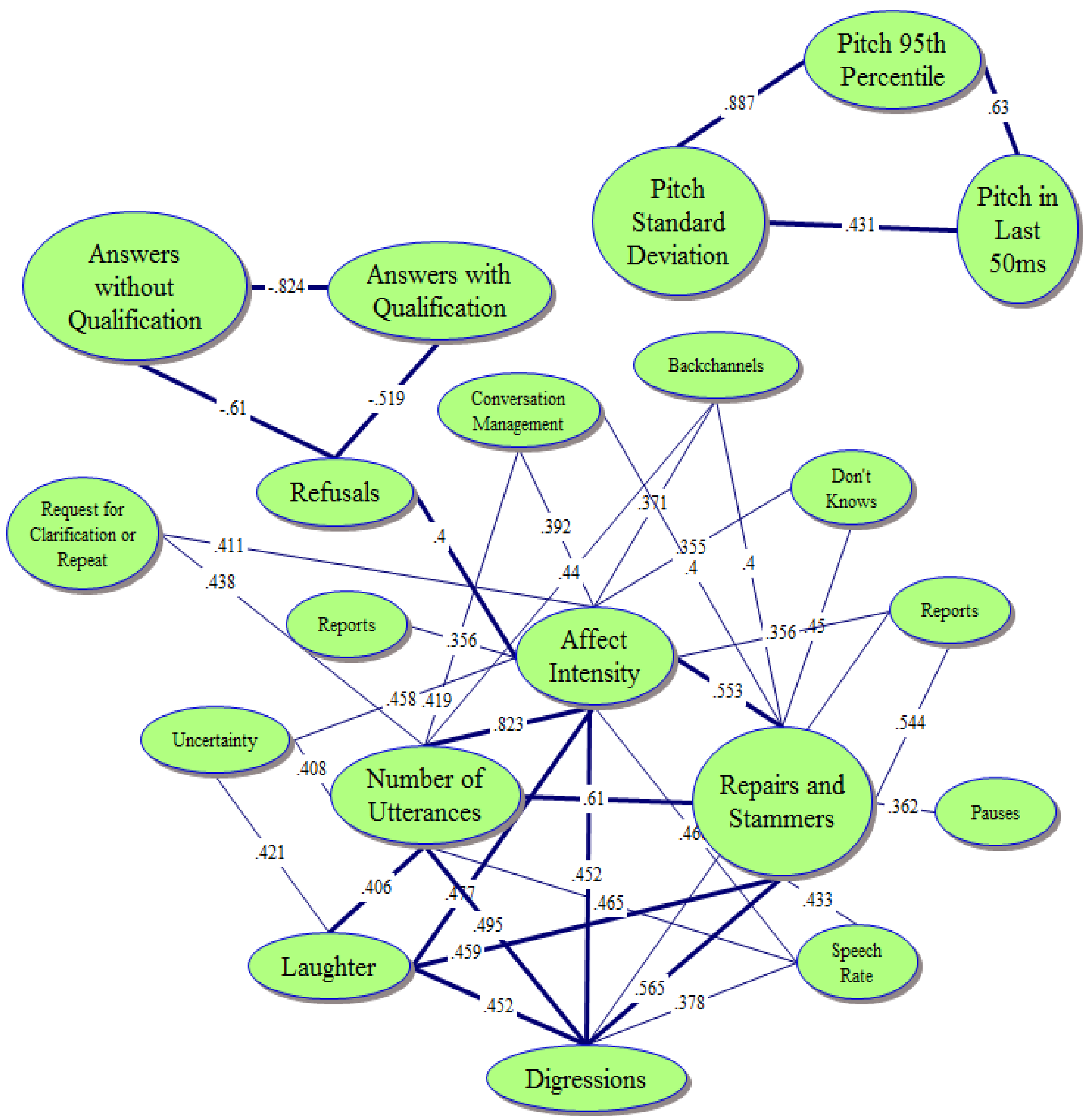
Income Nonresponse on…	**Estimate**	**SE**	**p-value**
Backchannel	**-0.902**	**0.385**	**0.019**
Conversation Management	0.120	0.430	0.780
Laughter	0.481	0.316	0.128
Affect Intensity	**-0.207**	**0.055**	**0.000**
Affect Valence	**-0.713**	**0.160**	**0.000**
Word Count	0.003	0.003	0.310
Overspeech	**0.069**	**0.025**	**0.006**
Median Pitch	0.000	0.003	0.868
Low Pitch	-0.002	0.004	0.560
High Pitch	0.002	0.002	0.204
Pitch SD	-0.007	0.006	0.204
Pitch Last 50ms	0.000	0.001	0.746
Speech Rate (syl/sec)	-0.041	0.058	0.483
Speech Rate (words)	0.045	0.069	0.521
Number of Utterances	-0.011	0.045	0.801
Duration of First Respondent Utterance	-0.027	0.024	0.257
Negative Comment	**2.081**	**1.107**	**0.060**
Answers without Qualification	**-1.109**	**0.298**	**0.000**
Answers with Qualification	**-0.849**	**0.285**	**0.003**
Request for Clarification or Repeat	0.251	0.254	0.324
No Difficulty	0.521	0.480	0.278
Report	**0.735**	**0.286**	**0.010**
Filler	-0.267	0.231	0.248
Pause	0.060	0.269	0.825
Refusal	**1.472**	**0.474**	**0.002**
Uncertainty	-0.080	0.337	0.812
Don't Know	**0.657**	**0.298**	**0.027**
Digression	-0.084	0.281	0.764
Repair and Stammer	0.252	0.248	0.309

Bracketed Response on…	**Estimate**	**SE**	**p-value**
Backchannel	-0.149	0.420	0.723
Conversation Management	-0.329	0.439	0.454
Laughter	0.062	0.297	0.835
Affect Intensity	0.058	0.042	0.165
Affect Valence	-0.109	0.159	0.493
Word Count	-0.004	0.004	0.282
Overspeech	-0.035	0.026	0.171
Median Pitch	0.000	0.003	0.892
Low Pitch	0.002	0.004	0.591
High Pitch	0.000	0.002	0.925
Pitch SD	-0.001	0.005	0.901
Pitch Last 50ms	0.002	0.001	0.087
Speech Rate (syl/sec)	0.003	0.055	0.955
Speech Rate (words)	0.040	0.068	0.557
Number of Utterances	0.021	0.047	0.657
Duration of First Respondent Utterance	-0.003	0.021	0.886
Negative Comment	-0.145	1.497	0.923
Answers without Qualification	-0.410	0.304	0.177
Answers with Qualification	-0.356	0.312	0.255
Request for Clarification or Repeat	0.069	0.268	0.796
No Difficulty	**-1.262**	**0.468**	**0.007**
Report	**0.714**	**0.275**	**0.010**
Filler	-0.265	0.235	0.259
Pause	0.026	0.259	0.920
Refusal	**1.685**	**0.476**	**0.000**
Uncertainty	-0.028	0.315	0.928
Don’t Know	-0.159	0.318	0.618
Digression	-0.261	0.272	0.337
Repair and Stammer	0.133	0.245	0.588

Appendix M: Correlogram of Most Highly Correlated Indicators

This figure presents the bivariate correlations between the most highly correlated indicators in the data set. The large number of moderate and high (r greater than .35) correlations is evident. It is also evident that several variables could logically be part of multiple factors, where clusters of three or more variables suggest factors. It is this overlapping of indicators across potential factors (i.e., high cross-loadings) that leads to problems with measurement model fit.

The variables with larger font and connected by thicker lines were used in the single factor score analyzed in this chapter.



References

Atrostic, B. K., & Kalenkoski, C. (2002). Item response rates: One indicator of how well we measure income. *Proceedings of the American Statistical Association, Survey Research Methods Section,* Washington, D.C. 94-99.

Bachorowski, J. A. (1999). Vocal expression and perception of emotion. *Current Directions in Psychological Science, 8*(2), 53-57.

Bachorowski, J. A., & Owren, M. J. (1999). Acoustic correlates of talker sex and individual talker identity are present in a short vowel segment produced in running speech. *The Journal of the Acoustical Society of America, 106*(2), 1054-1063.

Bargh, J. A., Chaiken, S., Govender, R., & Pratto, F. (1992). The generality of the automatic attitude activation effect. *J. of Personality and Social Psychology, 62*(6), 893-912.

Bargh, J. A., McKenna, K. Y. A., & Fitzsimons, G. M. (2002). Can you see the real me? Activation and expression of the "true self" on the internet. *Journal of Social Issues, 58*(1), 33-48.

Battaglia, M. P., Hoaglin, D. C., Izrael, D., Khare, M., & Mokdad, A. (2002). Improving income imputation by using partial income information and ecological variables. *Proceedings of the Joint Statistical Meetings,* New York City, 152-157.

Beatty, P., & Herrmann, D. (2002). To answer or not to answer: Decision processes related to survey item nonresponse. In R. M. Groves, D. A. Dillman, J. L. Eltinge & R. J. A. Little (Eds.), *Survey nonresponse* (pp. 71–85). New York: Wiley Series in Survey Methodology.

Bell, R. (1984). Item nonresponse in telephone surveys: An analysis of who fails to report income. *Social Science Quarterly, 65*(1), 207-215.

Benki, J. R. (2005a). "turnstats" Praat program for extracting turn information from digital audio recordings. Received Sept. 22, 2008

Benki, J. R. (2005b). "voicestats" Praat program for extracting turn information from digital audio recordings. Received Sept. 22, 2008

Bortfeld, H., Leon, S. D., Bloom, J. E., Schober, M. F., & Brennan, S. E. (2001). Disfluency rates in conversation: Effects of age, relationship, topic, role, and gender. *Language and Speech, 44*(2), 123-147.

Bradburn, N., Sudman, S., & Associates. (1979). *Improving interview method and questionnaire design.* San Francisco: Jossey-Bass.

U. S. Bureau of Labor Statistics, *Current population survey: Basic monthly survey questionnaire.* Retrieved December 22, 2009 from http://www.bls.census.gov/cps/bqestair.htm

Cacioppo, J. T., Petty, R. E., Feinstein, J. A., & Jarvis, W. B. G. (1996). Dispositional differences in cognitive motivation: The life and times of individuals varying in need for cognition. *Psychological Bulletin, 119*(2), 197-253.

Cannell, C. F., Fowler, Jr, F. J., & Marquis, K. H. (1968). The influence of interviewer and respondent psychological and behavioral variables on the reporting in household interviews. *Vital and Health Statistics, Series 2, Data Evaluation and Methods Research, 26*, 1-65.

Cassell, J., Gill, A. J., & Tepper, P. A. (2007) Coordination in conversation and rapport. *Proceedings of the Workshop on Embodied Language Processing,* Prague, 41-50.

Cassell, J., & Miller, P. (2008) Is it self-administration if the computer gives you encouraging looks. In F. G. Conrad and M. P. Schober (Eds.), *Envisioning the survey interview of the future* (pp. 161-178). New York: John Wiley & Sons.

Christianson, S. Å. (1992). *The handbook of emotion and memory: Research and theory.* Lawrence Erlbaum.

Conrad, F. G. & Schober, M. P. (2000). Clarifying question meaning in a household telephone survey. *Public Opinion Quarterly,* 64, 1-28

Conrad, F. G., Schober, M., & Dijkstra, W. (2008). Cues of communication difficulty in telephone interviews. In J. M. Lepkowski, et al. (Eds.), *Advances in telephone survey methodology* (pp. 212-230). New York: John Wiley & Sons.

Couper, M. P. (1998). Measuring survey quality in a CASIC environment. *Section on Survey Research Methods, Joint Statistical Meetings*, 41-49.

Couper, M. P., Singer, E., Conrad, F. G., & Groves, R. M. (2008). Risk of disclosure, perceptions of risk, and concerns about privacy and confidentiality as factors in survey participation. *Journal of Official Statistics*, *24*(2), 255-275.

Crowne, D. P., & Marlowe, D. (1960). A new scale of social desirability independent of psychopathology. *Journal of Consulting Psychology, 24*(4), 349-354.

de Leeuw, E. D. (1992). *Data quality in mail, telephone and face to face surveys.* Amsterdam: T. T. Publikaties.

de Leeuw, E. D., Hox, J., & Huisman, M. (2003). Prevention and treatment of item nonresponse. *Journal of Official Statistics, 19*(2), 153-176.

DePaulo, B. M., Lindsay, J. J., Malone, B. E., Muhlenbruck, L., Charlton, K., & Cooper, H. (2003). Cues to deception. *Psychological Bulletin, 129*(1), 74-118.

Diener, E., & Emmons, R. A. (1984). The independence of positive and negative affect. *Journal of Personality and Social Psychology, 47*(5), 1105-1117.

Dixon, J. (2005). Comparison of item and unit nonresponse in household surveys. *Washington DC, Bureau of Labor Statistics, Retrieved July, 2*, 2008.

Draisma, S., & Dijkstra, W. (2004). Response latency and (para) linguistic expressions as indicators of response error. In S. Presser, et al. (Eds.), *Methods for testing and evaluating survey questionnaires,* 131-148. Hoboken, NJ: John Wiley & Sons.

Dykema, J., Lepkowski, J. M., & Blixt, S. (1997). The effect of interviewer and respondent behavior on data quality: Analysis of interaction coding in a validation study. In L. Lyberg, P. Biemer, M. Collins, E. D. de Leeuw, C. Dippo, N. Schwarz, & D. Trewin (Eds.), *Survey measurement and process quality* (pp. 287–310). New York: John Wiley & Sons, Inc.

Ehlen, P., Schober, M. F., & Conrad, F. G. (2007). Modeling speech disfluency to predict conceptual misalignment in speech survey interfaces. *Discourse Processes, 44*(3), 245-265.

Ellsworth, P. C. (1994). William James and emotion: Is a century of fame worth a century of misunderstanding? *Psychological Review, 101*(2), 222-229.

Ericsson, K. A., & Simon, H. (1980). Verbal reports as data. *Psychological Review, 87*(3), 215-251.

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis*. MIT Press Cambridge, Mass.

Fabrigar, L. R., & Petty, R. E. (1999). The role of the affective and cognitive bases of attitudes in susceptibility to affectively and cognitively based persuasion. *Personality and Social Psychology Bulletin, 25*(3), 363-381.

Forgas, J. P. (2001). *Handbook of affect and social cognition.* Lawrence Erlbaum & Associates.

Fowler, F. J. (1995). *Improving survey questions: Design and evaluation.* Sage Publications, Inc.

Fowler, F. J., & Cannell, C. F. (1996). Using behavioral coding to identify cognitive problems with survey questions. In N. Schwarz & S. Sudman (Eds.), *Answering questions: Methodology for determining cognitive and communicative processes in survey research,* (pp. 15-36). Jossey-Bass.

Fromkin, V. (1973). *Speech errors as linguistic evidence*. The Hague: Mouton.

Gardner, H. (1999). *Intelligence reframed: Multiple intelligences for the 21st century* Basic Books.

Groves, R. M., Fowler, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2004). *Survey methodology.* Hoboken, NJ: Wiley-Interscience.

Groves, R. M., O'Hare, Barbara, C., Gould-Smith, D., Benki, J., & Maher, P. (2008). Telephone interviewer voice characteristics and the survey participation decision. In J. M. Lepkowski, et al. (Eds.), *Advances in telephone survey methodology* (pp. 385-400). New York: Wiley.

Groves, R. M. (2006). Nonresponse rates and nonresponse bias in household surveys. *Public Opinion Quarterly, 70*(5), 646.

Groves, R. M. (1989). *Survey errors and survey costs*. New York: Wiley.

Heeringa, S., Hill, D. H., & Howell, D. A. (1993). *Unfolding brackets for reducing item nonresponse in economic surveys,* AHEAD/HRS Report No. 94-029, Ann Arbor: Institute for Social Research

Hox, J., de Leeuw, E., & Vorst, H. (1996). A reasoned action explanation for survey nonresponse. In S. Laaksonen (Ed.), *International perspectives on nonresponse* (pp. 101-110). Helsinki: Statistics Finland.

Hurd, M., Juster, F. T., & Smith, J. P. (2003). Enhancing the quality of data on income: Recent innovations from the HRS. *The Journal of Human Resources, 38*(3), 758-772.

Juster, F. T., Smith, J. P., & Stafford, F. (1999). The measurement and structure of household wealth. *Labour Economics, 6*(2), 253-275.

Kahn, R. L., & Cannell, C. F. (1957). *The dynamics of interviewing; Theory, technique, and cases*. New York: Wiley.

Kent, R. D. (1997). *The speech sciences.* Singular.

Klatt, D. H., & Klatt, L. C. (1990). Analysis, synthesis, and perception of voice quality variations among female and male talkers. *The Journal of the Acoustical Society of America, 87*(2), 820-857.

Kormendi, E. (1988). The quality of income information in telephone and face to face surveys. In R. M. Groves, P. P. Biemer, L. E. Lyberg, J. T. Massey, W. L. Nichols, II, & J. Waksberg (Eds.), *Telephone survey methodology* (pp. 341-356). New York: Wiley.

Körmendi, E., & Noordhoek, J. (1989). *Data quality and telephone interviews.* Copenhagen, Denmark: Danmarks Statistik.

Leinonen, L., Hiltunen, T., Linnankoski, I., & Laakso, M. L. (1997). Expression of emotional–motivational connotations with a one-word utterance. *The Journal of the Acoustical Society of America, 102*, 1853.

Lewis, M., Haviland-Jones, J. M., & Barrett, L. F. (2000). *Handbook of emotions.* New York: Guilford Press.

Maynard, D. W., & Schaeffer, N. C. (2002a). *Standardization and tacit knowledge: Interaction and practice in the survey interview*. New York: Wiley.

Maynard, D. W. & Schaeffer, N. C. (2002b). Standardization and its discontents. In D. W. Maynard, H. Houtkoop-Steenstra, N. C. Schaeffer, & J. van der Zouwen, *Standardization and tacit knowledge: Interaction and practice in the survey interview* (pp 3-47). New York: John Wiley & Sons.

Maynard, D. W., Schaeffer, N. C., Drew, I. P., Raymond, G., & Weinberg, D. (2006). Standardization-in-interaction: The survey interview. In P. Drew, G. Raymond & D. Weinberg (Eds.), *Talk and interaction in social research methods* (pp. 9-27). London: Sage.

McGrath, D. (2005). Comparison of data obtained by telephone versus personal visit response in the US Consumer Expenditures Survey. *Joint Statistical Meetings of the American Statistical Association*, Minneapolis, MN.

Moore, J. C. (2006). The effects of questionnaire design changes on general income amount nonresponse in waves 1 and 2 of the 2004 SIPP panel. *Research Report Series, Survey Methodology, #2006-4,* Statistical Research Division, U.S. Census Bureau.

Moore, J. C., & Loomis, L. S. (2001). Using alternative question strategies to reduce income nonresponse. *Research Report Series (Survey Methodology)*, 2001 #03. Washington, DC: Statistical Research Division, U.S. Bureau of the Census.

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

Nicoletti, C., & Peracchi, F. (2001). *Sample participation and income nonresponse in the ECHP*. Unpublished manuscript.

Nisbett, R. E. & Wilson, T. D. (2005). Telling more than we can know: Verbal reports on mental processes. In D. L. Hamilton (ed.) *Key readings in social psychology: Social cognition,* pp. 210-227. New York: Psychology Press.

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

Olson, L., Rodén, S., Dennis, M., & Cannarozzi, F. (1999). Alternative methods of obtaining family income in RDD surveys. *American Association for Public Opinion Research Annual Conference,* 940-945, St. Pete Beach, FL.

Ongena, Y. P., & Dijkstra, W. (2007). A model of cognitive processes and conversational principles in survey interview interaction. *Applied Cognitive Psychology, 21*(2), 145-163.

Philippot, P., Feldman, R. S., & Coats, E. J.,. (1999). *The social context of nonverbal behavior*. Cambridge, UK: Cambridge University Press; Editions de la Maison des sciences de l'homme.

Riphahn, R. T., & Serfling, O. (2005). Item non-response on income and wealth questions. *Empirical Economics, 30*(2), 521-538.

Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology, 39*(6), 1161-1178.

Schachter, S., & Singer, J. (1964). The interaction of cognitive and physiological determinants of emotional state. *Advances in Experimental Social Psychology, 1*, 49-80.

Schaeffer, N. C. (2002). Conversation with a purpose–or conversation? Interaction in the standardized survey interview. In D. W. Maynard, H. Houtkoop-Steenstra, N. C. Schaeffer, & J. van der Zouwen (Eds.), *Standardization and tacit knowledge: Interaction and practice in the survey interview* (pp. 95–123). New York: John Wiley & Sons, Inc.

Schaeffer, N. C. & Maynard, D. W. (2002). Occasions for intervention: Interactional resources for comprehension in standardized survey interviews. In D. W. Maynard, H. Houtkoop-Steenstra, N. C. Schaeffer, & J. van der Zouwen, (Eds.), *Standardization and tacit knowledge: Interaction and practice in the survey interview* (pp 261-281). New York: John Wiley & Sons.

Schaeffer, N. C. & Maynard, D. W. (2007). The Contemporary standardized survey interview for social research. In F. G. Conrad & M. F. Schober (Eds.), *Envisioning the survey interview of the future* (pp. 31-57). New York: Wiley.

Schober, M. P., & Conrad, F. G. (1997). Does conversational interviewing reduce survey measurement error? *Public Opinion Quarterly, 61*(4), *576-602.*

Schober, M. F., & Bloom, J. E. (2004). Discourse cues that respondents have misunderstood survey questions. *Discourse Processes, 38*(3), 287-308.

Schraepler, J. P., Schupp, J., & Wagner, G. (2006). Changing from PAPI to CAPI: A longitudinal study of mode-effects based on an experimental design. *Paper Presented at the 2006 European Conference on Quality in Survey Statistics*, Cardiff, UK.

Schuman, H. & Presser, S. (1977). Question wording as an independent variable in survey analysis. *Sociological Methods & Research, 6*(2), 151-170.

Schwarz, N. (2000). Emotion, cognition, and decision making. *Cognition and Emotion, 14*(4), 433-440.

Schwarz, N., & Clore, G. L. (2007). Feelings and emotional experiences. In A. W. Kruglanski & E. T. Higgins (Eds.), *Social psychology: Handbook of basic principles* (pp. 385-407). Guilford Press.

Schwarz, N., & Clore, G. L. (2003). Mood as information: 20 years later. *Psychological Inquiry, 14*(3 & 4), 296-303.

Schwarz, N. (2000). *Emotion, cognition, and decision making*. East Sussex, UK: Psychology Press.

Sirken, M. G., Herrmann, D. J., Schechter, S., Schwarz, N., Tanur, J. M., & Tourangeau, R. (Eds). (1999). *Cognition and survey research.* New York: John Wiley & Sons, Inc.

Singer, E., Mathiowetz, N. A., & Couper, M. P. (1993). The Impact of privacy and confidentiality concerns on survey participation: The Case of the 1990 Census. *The Public Opinion Quarterly*, 57(4), 465-482.

Sternberg, R. J., Lautrey, J., & Lubart, T. I. (2003). Models of intelligence: International perspectives (pp. 373). Washington, DC: American Psychological Association.

Suchman, L. & Jordan, B. (1990). Interactional troubles in face-to-face survey interviews. *Journal of the American Statistical Association, 85*(409), 232-241.

Suchman, L. & Jordan, B. (1991). Validity and the collaborative construction of meaning in face-to-face surveys. In J. M. Tanur (Ed.), *Questions about questions: Inquiries into the cognitive bases of surveys* (pp. 241–67). New York: Sage.

Thomas, E. A. & Malone, T. W. (1979). On the dynamics of two-person interaction. *Psychological Review, 86*(4), 331-360.

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The psychology of survey response* Cambridge University Press.

Tourangeau, R., & Smith, T. W. (1996). Asking sensitive questions: The impact of data collection mode, question format, and question context. *Public Opinion Quarterly, 60*(2), 275-304.

Tourangeau, R., Smith, T. W., & Rasinski, K. A. (1997). Motivation to report sensitive behaviors on surveys: Evidence from a bogus pipeline experiment. *Journal of Applied Social Psychology, 27*(3), 209-222.

Tourangeau, R., & Yan, T. (2007). Sensitive questions in surveys. *Psychological Bulletin, 133*(5), 859-883.

Traugott, M. (2008). Validation studies. In W. Donsbach & M. Traugott (Eds), *The Sage handbook of public opinion research* (pp. 408-416). New York: Sage.

Van der Vaart, W., Ongena, Y., Hoogendoorn, A., & Dijkstra, W. (2006). Do interviewers' voice characteristics influence cooperation rates in telephone surveys? *International Journal of Public Opinion Research, 18*(4), 488.

Watt, J. H. & vanLear, C. A. (1996). *Dynamic patterns in communication processes*. New York: Sage.

Webb, E. J. (2000). *Unobtrusive measures.* New York: Sage Publications Inc.

Webb, E. J., Campbell, D. T., Schwartz, R. D., & Sechrest, L. (1966). Unobtrusive measures: Nonreactive research in the social sciences. *PROJECT C-998*, 237.

Willis, G. B. (2005). *Cognitive interviewing: A Tool for improving questionnaire design.* Thousand Oaks, CA: Sage Publications, Inc.

Yan, T., Jans, M., & Curtin, R. (2006). Changes in nonresponse to income questions. *Proceedingss Joint Statistcal Meetings/American Association for Public Opinion Research,* Montreal, Canada.

Zajonc, R. B. (1980). Feeling and thinking: Preferences need no inferences. *American Psychologist, 35*(2), 151-175.