2013

Collecting Paradata for Measurement Error Evaluations

Kristen Olson
University of Nebraska - Lincoln, kolson5@unl.edu

Bryan Parkhurst
University of Nebraska–Lincoln, bryanparkhurst@unl.edu

Published as Chapter 3 in *Improving Surveys with Paradata: Analytic Uses of Process Information,* First Edition. Edited by Frauke Kreuter. John Wiley & Sons, 2013, pp. 43–72.

Collecting Paradata for Measurement Error Evaluations

Kristen Olson and Bryan Parkhurst

University of Nebraska–Lincoln

1 Introduction

Survey researchers and methodologists seek to have new and innovative ways of evaluating the quality of data collected from sample surveys. Paradata, or data collected for free from computerized survey instruments, have increasingly been used in survey methodological work for this purpose (Couper, 1998). One error source that has been studied using paradata is measurement error, or the deviation of a response from a "true" value (Groves, 1989; Biemer and Lyberg, 2003). Although used in psychological literature since the 1980s (see Fazio, 1990, for an early review) and adapted to telephone interviews by Bassili in the early 1990s (Bassili and Fletcher, 1991; Bassili and Scott, 1996), the adoption and use of paradata for studying measurement-error- related outcomes has grown exponentially with the growth of web surveys and increased use of computerization in interviewer-administered surveys (Couper, 1998; Heerwegh, 2003; Couper and Lyberg, 2005). Paradata are a proxy for breakdowns in the cognitive response process or identify problems respondents and interviewers have with a survey instrument (Couper, 2000; Yan and Tourangeau, 2008).

Paradata can be collected at a variety of levels, resulting in a complex, hierarchical data structure. Examples of paradata collected automatically by many computerized survey software systems include timing data, keystroke data, mouse click data, and information about the type of interface such as the web browser and screen resolution. Examples of paradata that inform the measurement process, but not collected automatically, include behavior codes, analysis of vocal characteristics, and interviewer evaluations or observations of the survey-taking process. Paradata available to be captured vary by mode of data collection and the software used for data collection. One challenge is that not all off-the-shelf software programs capture paradata, and thus user-generated programs

have been developed to assist in recording paradata. Further complicating matters is how the data are recorded, ranging from text or sound files to ready-to-analyze variables. **In** this chapter, we review different types of paradata, evaluate how paradata differs by mode, and examine how to turn paradata into an analytic dataset. This chapter does not review paradata kept about the recruitment effort, including number of call attempts, indicators of refusals, or observations of a sampled housing unit. For a discussion of these types of paradata, see Chapter 2 (of *Improving Surveys with Paradata: Analytic Uses of Process Information,* ed. Frauke Kreuter, John Wiley & Sons, 2013).

2 Paradata and Measurement Error

Measurement error occurs when a respondent's answer differs from a conceptual "true value." These deviations between answers and "truth" occur when there is a breakdown in the cognitive response process (Tourangeau et al., 2000; Dillman et al., 2009). As shown in Figure 1, the cognitive response process consists of five general steps (four in interviewer-administered questions). First, perception involves seeing the graphical layout and images of a self-administered survey. Second, comprehension involves understanding the words and concepts being asked about in a survey question and response options. Third, retrieval is the process of recalling or generating the relevant material from memory. Next, judgment involves mapping the retrieved information onto the response options or response format. Finally, editing involves changing the retrieved and mapped information when responding to a question in response to social desirability, sensitivity, or privacy concerns. If a breakdown occurs at any of these stages, then the response that ends up in the final dataset will not reflect "truth" or the question may not be answered at all (Beatty and Herrmann, 2002; Krosnick, 2002). If breakdowns of the cognitive response process occur systematically in the same direction over all respondents then a measurement error bias will result. If these breakdowns occur with varying magnitude and direction across respondents, then a measurement error variance will result.

As shown in Table 1, paradata have been used to detect a wide variety of breakdowns of the cognitive response process using both observational (indicated by (S) for survey) and experimental methods (indicated by (E) for experiment). Paradata such as behavior codes and interviewer evaluations are somewhat more easily interpreted because they measure more concretely interpreted constructs (e.g., question read exactly as written, rating of the respondent's cooperativeness) and, as such, are excluded from Table 1. Using response latencies as an indicator

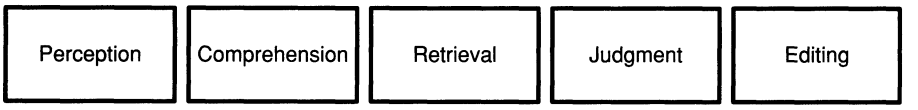

Figure 1. Cognitive response process. From Dillman et al. (2009) and Tourangeau et al. (2000).

Table 1. Examples of Operationaiization of Paradata for Measurement Error Studies

Paradata	Operationalization	Interpretation	Example Studies
Response latencies	Long response time	Ambivalent attitudes	Bassili & Fletcher, 1991 (S); Mulligan et al., 2003 (S); Johnson, 2004 (S); Heerwegh, 2003 (S)
		Lack of knowledge	Heerwegh, 2003 (S)
		Poor question wording	Bassili, 1996 (S); Bassili & Scott, 1996 (S); Yan & Tourangeau, 2008 (S)
		Longer cognitive processing	Yan & Tourangeau, 2008 (S); Stieger & Reips, 2010 (S); Malhotra, 2008 (E)
		Complex visual layouts	Stern, 2008 (E); Healey, 2007 (E), Heerwegh, 2002 (E)
		Interviewer building rapport	Penne et al., 2002 (S)
		Inconsistent response options	Tourangeau et al., 2004 (E)
		Usability problems	McClamroch, 2011 (S); Bassili & Scott, 1996 (S); Healey, 2007 (E); Heerwegh & Loosveldt, 2002 (E)
		Engaged respondent	Crawford, Couper & Lamias, 2001 (E)
	Short response time	Accessible attitudes	Bassili & Fletcher, 1991 (S); Mulligan, et al., 2003 (S); Johnson, 2004 (S); Grant et al., 2010 (S)
		Knowledge	Heerwegh, 2003 (S); Fazio, 1990 (E)
		Logical question order	Tourangeau, et al., 2004 (E)
		Ease of use for open-ended questions compared to search through long list	Stern, 2008 (E)
		Interviewer falsification or shortcutting	Johnson et al., 2001 (nJa); Penne et al., 2002 (S)
		“Speeders”	Roßman, 2010 (S); Gutierrez, Wells, Rao, & Kurzynski, 2011 (S); Stieger & Reips, 2010 (S)
		More working memory capacity	Yan and Tourangeau, 2008 (S)
		Expertise with mode	Yan and Tourangeau, 2008 (S)
		Answering before question is fully read	Caspar & Couper, 1997 (E)

(continued)

Table 1. (continued)

Paradata	Operationalization	Interpretation	Example Studies
Mouse clicks	Answer changes	Guessing on knowledge questions	Heerwegh, 2003 (S)
		Uncertainty on attitude questions *or* optimizing behavior of reconsidering answers	Heerwegh, 2003 (S)
		Confusion in mapping response option verbal and numerical labels	Stern, 2008 (E); Heerwegh, 2011 (E)
		Subtraction effects	Stern, 2008 (E)
		Difficulty searching a long list	Stern, 2008 (E)
		Usability issues with scroll mice	Healey, 2007 (E)
		Reconstruct failures	Ostergren & Liu, 2010 (n/a)
Keystrokes	Answer changes	Incorrect responses to sensitive questions	Penne et al., 2002 (S)
	Backing up	Interviewer browsing through questionnaire	Sperry, et al., 1998 (S)

(S) indicates that the study used observational survey data with no experimental variation. (E) indicates that the study used experimental data from a laboratory or field setting. n/a indicates that the article mentions the use of paradata for this purpose, but does not empirically examine paradata.

of accessible versus ambivalent attitudes is one of the most common uses of this form of paradata (Bassili and Fletcher, 1991; Mulligan et al., 2003; Johnson, 2004). Short response latencies indicate accessible attitudes and long response latencies indicate more ambivalent attitudes. Yet, response latencies have been used for a variety of other problems encountered when answering questions including lack of knowledge (an encoding issue or poor retrieval), poor question wording (comprehension), confusion in the meaning of numeric versus verbal response option labels (judgment), and issues related to question-answering external to the cognitive response process such as an interviewer building rapport with the respondent or a respondent's familiarity with a mode. In fact, the panoply of research illustrated by Table 1 shows that paradata in general (and especially response latencies in particular) have been used to reflect all stages of the cognitive response process. To make the relevance of each stage of the cognitive response process explicit, Johnson (2004, p. 685-687) and Chessa and Holleman (2007) have incorporated multiple cognitive steps into mathematical models for response latencies. We now turn to a more detailed description of each type of paradata.

3 Types of Paradata

There are many types of paradata collected "for free" by computerized survey software programs. By "for free," we mean that the paradata do not necessarily cost additional interviewer hours to collect or add to respondent burden, but they may require additional programming and data storage costs depending on the software being used. These paradata include time stamps, keystroke files, mouse click files, and digital audio recordings (Kreuter and Casas-Cordero, 2010; Heerwegh, 2011b). Measurement-error-related paradata can be collected at four levels of aggregation: the survey level, the section level, the question level, and the action level (Kaczmirek, 2008; Heerwegh, 2011b). Additionally, there are other data about the measurement process that have been used for purposes of measurement error evaluation, although not necessarily captured "for free" from a computerized survey software system. As with the paradata collected for free, these types of paradata can also be captured at various levels of aggregation. These include behavior codes, analysis of vocal characteristics, and interviewer evaluations of the survey-taking environment. Although eye tracking methods are increasingly being used to understand how respondents process and respond to a self-administered questionnaire, as of this writing, these technologies are only used in a lab setting and not in field production. Thus, they are excluded from this chapter, but interested readers are referred to Galesic et al. (2008) and Galesic and Yan (2011) for overviews.

3.1 Time Stamps

The most common type of paradata collected is time stamps. Time stamps record the date and time when actions occur in a survey. An action can be anything from viewing the first screen of a survey to entering the value of "2" in a response

field. Time stamps may be recorded concurrently with other actions such as keystrokes and mouse clicks (described below), but may simply be recorded at the onset or offset of individual screens of a computerized questionnaire. These actions can be recorded at four different levels of aggregation and refinement, from the start and end time of an interview that is recorded to the minute to recording when each keystroke or mouse click in a survey occurs to the millisecond (1/1000 of 1 s).

At the most aggregate level, survey-level time stamps record the date and time that a survey is initiated and completed. The difference between the survey's end time and start time is the length of the survey, usually reported in minutes. Survey-level time stamps can be recorded by interviewers with a wristwatch, although in today's surveys, most computer-assisted survey interviewing programs use internal clocks to record the time it takes to complete a questionnaire or interview. Length of interview has long been standard practice to record and include in datasets—for example, the 1959 Detroit Area Studies survey includes the length of interview in its public use dataset (Swanson and Brazer, 1959). The total length of interview can be used for a variety of purposes. Records of the length of time for each interview are recommended to be kept in the ISO Standards 20252 for purposes of interview verification (International Organization for Standardization, 2006). In addition, the total length of interview has been to understand interviewer behaviors (Olson and Peytchev, 2007; Olson and Bilgen, 2011) and as a measure of respondent commitment (Malhotra, 2008). .

At a lower level of aggregation are section-level timings, in which interviewers or the computer's internal clock records the start and end date and time of each section of the survey. Differences between the time of initiation and completion for each section of the survey provide a measure of the length of the section, also usually reported in minutes. Although sections are often transparent to a respondent, paradata measuring section timings provide information on how long predefined blocks of questions take to complete. Section timings have been used to evaluate respondent fatigue in a Audio Computer-Assisted Self-Interview (ACASI) (Caspar and Couper, 1997), to examine cross-national differences in questionnaire length (Jurges, 2005), and to compare with the same sections using paper and pencil instruments (Burrell, 2003), for example. Section timings that are unusually short or unusually long signal to the researcher that the respondent or interviewer did not take the appropriate amount of time with the section or had unusual difficulties.

Question-level timings measure how long each question takes to administer and answer. Paradata recorded automatically by a computer from the time of the first display of the question to the time of advancing to the next question includes more than simply respondent cognitive processing time. In interviewer-administered surveys, this includes the time it takes the interviewer to read the question, the time between the end of the question reading and the respondent's answer, and the actual length of time of the respondent's answer, including any questions, clarifications and other verbal utterances (Bassili, 1993; Mulligan et al., 2003). In web surveys, server-side timings of questions include the time to download the question from the server, be displayed on the respondent's computer, for the respondent to answer, and for the answer to be transmitted back to the server (Yan and Tourangeau, 2008). For example, as shown in Figure 2, a CATI software

```
5 F1A 7.44 WED OCT 19 08:13:21 2011
3 F1B 7.50 WED OCT 19 08:13:23 2011
```

Figure 2. Example of question level timing data from a CATI software system.

program may export question-level paradata as a series of rows with all cases aggregated into one text file.

In interviewer-administered surveys, alternative methods have been used to measure question-level timings. Some studies have refined question-level timing using interviewers to identify the time elapsed between the reading of a question and the onset of the respondent's answer (Bassili and Fletcher, 1991; Bassili, 1993; Johnson, 2004); these studies, however, move beyond the data that can be captured "for free" and turn the interviewer into the paradata-collecting agent. Still others use clocks triggered by vocal utterances (Bassili and Fletcher, 1991; Bassili, 1993). Question-level timing data have been used to understand cognitive processes (Smyth et al., 2006; Yan and Tourangeau, 2008), visual processes (Tourangeau et al., 2004; Stern, 2008), and to trigger tailored communication to the respondent (Conrad et al., 2007). For example, Smyth et al. (2006) found that respondents spent longer times on each question when asked in a forced choice format than a check-all-that-apply format, a finding that they attribute to deeper cognitive processing for forced choice questions.

Finally, the most micro-level collection of time stamp paradata occurs at the keystroke level or mouse click level. In these "action-level" timing paradata, time is recorded for each action taken by the person interacting with the computerized questionnaire, either the interviewer or the respondent. Action-level paradata are complicated because, unlike the number of sections or number of questions in a survey, the number of actions taken by the respondents or the interviewer is not fixed in advance and varies across respondents. As such, these data are non-rectangular, that is, there are varying numbers of observations per respondent. However, these data are also the richest in terms of insights into what is going wrong in a survey instrument. Since response timing data at the action level are most frequently used in conjunction with the keys or mouse clicks that generate them, we will discuss these timing paradata in the next two sections.

3.2 Keystrokes

Keystroke files, sometimes called audit trails or trace files, are a second form of paradata, recording when interviewers or respondents used specific keys during the survey. That is, keystroke files contain both response timing data and a record of the keystrokes pressed during the questionnaire administration. Keystroke files are important because, in addition to recording when actions occur, they record which actions occur during the survey, allowing researchers to identify areas of difficulty for the interviewer or the respondent (Couper, 1998). Often of interest are the use of keystrokes for optional special function keys, help menus, backspaces, and Page Up/Page Down. As with timing data, keystroke files are often recorded at a micro-level (e.g., for each keystroke made), but can also be

recorded at the question level or aggregated to a section level or survey level. Unlike timing information, keystroke files are looking for the presence or absence of a certain key being pressed rather than solely recording the time of an event (of course, keystroke files often record when the action occurred as well as what occurred). Aggregation is relatively straightforward: an action-level keystroke indicator (e.g., whether or not a help menu was accessed) can be summed up to the relevant question level (e.g., the total number of times a help menu was accessed for the question), which in turn can be summed to a measure of keystrokes within sections (e.g., the total number of times a help menu was accessed during Section A) or over entire questionnaire (e.g., the total number of times a help menu was accessed during the entire survey).

Keystroke data are potentially the richest source of paradata for understanding usability of questionnaires for interviewers (Couper et al., 1997; Sperry et al., 1998), identifying problem questions (Hansen and Marvin, 2001), and revealing whether the audio track is listened to in an ACASI interview (Caspar and Couper, 1997; Bart, 2007; Couper et al., 2009), among many other uses. For example, Couper et al. (1997) found that 92% of all interviews for the Assets and Health Dynamics Among the Oldest Old (AHEAD) study pressed a "comments" key (represented by function key F2) at least once during an interview, and that the use of this key declined as interviewer experience increased. Hansen and Marvin (2001) used keystrokes to identify "abnormal terminations" of a National Survey of Family Growth interview, that is, an interviewer stopping an interview at a question that was not the last screen of the CAPI instrument. They found an unusually high rate of exits (8.5%) at a question about pregnancy outcomes, and, after discussions with the interviewers, identified that the CAPI software took an unusually long time to record answers on that screen. Keystroke files are most often seen in interviewer-administered surveys, rather than web surveys, due to the keyboard-driven interface of most of the survey software for telephone and in-person modes.

3.3 Mouse Clicks

Mouse click files record each action the respondent or interviewer takes using the computer's mouse, ranging from the presence or absence of simple single mouse clicks to the position of the mouse cursor at a specified time interval on an x - y coordinate of the survey page (Heerwegh, 2003; Kaczmirek, 2008; Stieger and Reips, 2010; Heerwegh, 2011b). In a web survey, recordings of mouse clicks are called "client-side paradata" (CSP) (Heerwegh, 2003, 2011b). Simple JavaScript allows for easy and unobtrusive collection of a variety of actions, including entering answers using radio buttons, drop-down menus, or text fields, clicking in the "wrong" place, changing answers, and mouse movements, all collected invisibly to the respondent. For example, Dirk Heerwegh's JavaScript outputs files shown in Figure 3.

CSP is distinct from "server-side paradata," which is the information routinely recorded by survey software such as survey webpage submission dates and times, but does not include what happens on the respondent's side of the survey. A useful feature of CSP is that it allows researchers to gain "information about how respondents construct answers in their natural setting" (Stern, 2008, p. 379).

```
Sun Oct 23 2011 15:30:34 GMT-0500 (Central Daylight Time)
ft_fullDownload=3 ft=244:WindowFocusft=272:StartScrollft=498:EndScroll
at 1475pxft=965:v[1]=2ft=2541:v[1]=4ft=1566:v[1]=1ft=1594:target=#v2f
t=337:StartScrollft=503:EndScroll at 3006pxft=630:v[2 1]=clicked
ft=1580:v[2 1]=2ft=2144:v[2 1]=3ft=770:v[2 2]=clickedft=812:v[2 2]=1ft=
2031:target=#v3ft=30:StartScrollft=500:EndScroll at 4410pxf
t=1490:v[3]=7ft=1106:v[3]=4ft=1220:target=#v4ft=183:StartScrollft=500:
EndScroll at 5875pxft=672:v[4]=6 ft=6705:v[4b]=something fabulousf
t=84:v[4]=77ft=1091:v[4]=6ft=1024:target=#v5 ft=425:StartScroll
ft=500:EndScroll at 7416pxft=7364:v[5]=changedft=11:WindowBlurf
t=1654:WindowFocusft=1576:form submittedf
```

Figure 3. Example of CSP from a web survey from Heerwegh's CSP project (https://perswww.kuleuven.bel~u0034437/public/csp.htm).

Generally used in web surveys, mouse click files have been used to examine choosing answer boxes, radio button, and list boxes, selecting a hyperlink, selecting an answer choice from drop-down menus, and mouse clicks on the submission buttons of the survey (Heerwegh, 2003; Peytchev et al., 2006; Conrad et al., 2006; Healey, 2007; Heerwegh, 2011b). This source of paradata can aid researchers in identifying various problem questions or sections where certain actions were more likely to occur (e.g., more answers changed) as potential indicators of lower data quality (Stern, 2008).

3.4 Behavior Codes

Behavior codes are information about the interviewer and respondent's verbal behaviors during a survey interview's question-answer process. They are developed and recorded by human coders, not automatically coded by computers. To obtain behavior codes, interviews are audio recorded (generally digitally today, but cassette tapes have been used in the past), transcribed, and then coded by a set of at least two coders to identify relevant behaviors. Alternatively, interviews can be monitored in real time and actions by the interviewer and respondent recorded while the interview is being conducted. Behavior codes may be recorded at aggregate levels such as a question or at smaller levels such as a "turn" or "utterance," that is, (portions of) each individual statement made by each actor within each question. Although many different types of behaviors have been coded, consistently used behavior codes are those related to the survey task itself, such as interviewer behaviors including reading a question exactly as written, probing behaviors, and providing feedback and respondent behaviors including providing an adequate answer, asking for clarification, and expressing uncertainty about an answer (Mathiowetz and Cannell, 1980; Schaeffer, 1991; Dykema et al., 1997; Fowler, 2011). Table 2 provides an example of six interviewers and six respondent behaviors coded in the Health Field Study (Belli et al., 2001a).

Behavior codes are not collected "for free" because human coders are used rather than computers. To reduce costs, behavior codes are often recorded for a subsample (random or convenience, depending on the goals) of the entire respondent pool. Additionally, since human coders are used, behavior codes have their

Table 2. Example of Behavior Codes from Belli et al. (1999, p. 198)

Interviewer Codes	Respondent Codes
Q-E Exact: Reads exactly as written or makes insignificant changes	**R-I** Interruption: Interrupts question with an answer.
Q-S Significant changes: Makes wording changes that can affect written question meaning.	**R-C** Clarification: Expresses uncertainty, requests question repetition, or seeks clarification.
Q-O Other changes: Verifies, states, or suggests an answer; reads inapplicable question; skips applicable question.	**R-Q** Qualified response: Qualifies answer with phrases such as "about," "I guess," "maybe," etc.
P-A Adequate probing: Probing is nondirective and sufficient.	**R-CR** Respondent corrects a response to a previous question.
P-D Directive probing: At least one probe is directive.	**R-D** Respondent digresses.
I-D Interviewer introduces digression: Digressions are verbal comments that are not directly related to satisfying question objectives.	**R-L** Respondent laughs.

own measurement error properties, with intracoder reliability often measured using kappa statistics; unreliably coded behaviors can be excluded from analyses. Unreliably coded behaviors often have a kappa value less than 0.40 (Bilgen and Belli, 2010), following recommendations from Landis and Koch (1977) and Fleiss et al. (2004). Although measurement error models have been developed to account for known measures of unreliability (Fuller, 1987), to our knowledge, these types of models have not been directly applied to analyses using behavior codes.

Fowler (2011) and Schaeffer and Dykema (2011) examine a breadth of literature on the use and relationship between particular behavior codes and measurement quality. Fowler (2011) summarizes three types of behavior coding studies, "link[ing] observed behaviors to the characteristics of questions, ... observed behaviors to interviewer-related error," and "observed behaviors to the "validity" of estimates from surveys" (p. 15). Behavior codes have been used as a pretesting method (Presser and Blair, 1994), to monitor interviewers (Mathiowetz and Cannell, 1980), to identify poorly written questions (Fowler, 2011), to get insights into the interaction that interviewers and respondents have during field interviews (Suchman and Jordan, 1990; Schaeffer, 1991), and as correlates of measurement error bias (Dykema et al., 1997; Mathiowetz, 1998; Belli et al., 2004) and interviewer-related variance (Mangione et al., 1992).

These behavior codes have been used in concert with other measures of paradata, such as response latencies (Draisma and Dijkstra, 2004) as joint predictors of measurement error. For example, Garbarski et al. (2011) examine the relationship between behaviors during a survey interview, response latencies, and responses to the self-reported health question in the Wisconsin Longitudinal Survey. They find support for their hypothesis that behaviors indicating problems during the question-answer sequence and longer response latencies are associated with worse self-reported health due to increased response task complexity (proxied by

the respondent behaviors) and decreased cognitive abilities (proxied by response latency) for those with worse health. In a meta-analytic approach, Mangione et al. (1992) examine the association between the prevalence of interviewer behaviors on particular questions and estimates of interviewer variance (measured with an intraclass correlation coefficient), finding higher levels of interviewer variance for questions that require more probing by interviewers. Although labor intensive, behavior codes permit survey researchers and methodologists insights into how the interaction between the respondent and interviewer can affect measurement error in survey questions.

3.5 Vocal Characteristics

Analysis of vocal characteristics, also called paralinguistic data (Draisma and Dijkstra, 2004), like behavior codes, examines audio recordings of interviews to identify notable traits of the interviewer's voice itself, rather than behaviors during the interview. These vocal properties include pitch (higher or lower sounding voices), intonation (rising or falling pitch), speech rate, and loudness (Oksenberg et al., 1986; Bachorowski, 1999; Bänziger and Scherer, 2005; Jans, 2010). These vocal properties are obtained by sending sound files through a computerized analysis program (such as Praat, http://www.praat.org). Vocal characteristics can be coded for a single word or sound (phoneme) ("Hello"), for phrases or for sentences. As with other forms of paradata, aggregation of vocal characteristics is used. Measures of central tendency (mean pitch) and variability (standard deviation of pitch, range of pitch) can be calculated for each sound file. Depending on the analyst's decision, these measurements can be at a word level, turn level, question level, or section level.

Interestingly, unit nonresponse rates have been examined as outcomes for vocal properties (Oksenberg et al., 1986; Oksenberg and Cannell 1988; van der Vaart et al., 2005; Groves et al., 2008; Benki et al., 2011 and Kreuter and Olson, Chapter 2) but little research exists examining the relationship between vocal properties and measurement error in survey questions. Two studies have examined the relationship between an interviewer's rising versus falling intonation on "yes-no" questions and acquiescence, finding contradictory results (Barath and Cannell, 1976; Blair, 1977). A third study has examined item nonresponse, a failure of the measurement process, and found no clear evidence of an association between pitch and item nonresponse on income questions (Jans, 2010). With these few studies, the realm of research areas for vocal characteristics paradata and measurement error is wide open.

An additional set of vocal characteristics distinct from those typically examined through behavior codes are interruptions to a fluid speech pattern, such as disfluencies ("uh," "um;" Ehlen et al., 2007), backchannels ("I see," "uh huh;" Conrad et al., 2013; Jans, 2010), or laughter (Bilgen, 2011). These behaviors are not directly task related, but instead are related to normal conversational behaviors (Jans, 2010). Disfluencies in survey interviews, unlike other vocal characteristics, have been shown to be related to comprehension problems (Schober and Bloom, 2004; Ehlen et al., 2007), difficulties with cognitive ability tasks (Schaeffer et al., 2008), but not to item nonresponse (Jans, 2010). For example, in a lab study,

Schober and Bloom (2004) found that scenarios with "complicated" mappings, that is, where the lab stimulus did not neatly map into an answer for a survey question, yield more "uhs" and "ums" than those with "straightforward" mappings in which the stimulus and survey question were more easily aligned. These few studies suggest that disfluencies may be a rich source of verbal paradata for future research on measurement errors.

3.6 Interviewer Evaluations

In interviewer-administered surveys, interviewers have long been asked to make general assessments about how engaged, cooperative, hostile, or attentive the respondent was during the interview (Feldman et al., 1951). Additionally, interviewers record information about the interview-taking environment, such as whether other individuals were present or whether the respondent used headphones during an ACASI component (Couper et al., 2009). Unlike the previous sources of paradata, these interviewer evaluations are questions asked directly of the interviewer and included as a few additional questions in the questionnaire. For example, the General Social Survey asks interviewers, "In general, what was the respondent's attitude toward the interview?" with response options "friendly and interested," "cooperative but not particularly interested," "impatient and restless," and "hostile" (Davis et al., 2007, p. 318). Also unlike the previous types of paradata, these sets of evaluations are almost always made solely for an entire survey, although occasionally observations for particular sections (e.g., sensitive questions, ACASI components) will be made.

As with response latencies, interviewer evaluations have a wide variety of applications. Interviewer evaluations have been used as proxies for rapport (Goudy and Potter, 1975), interviewer motivation (Olson and Peytchev, 2007), measures of the quality of the interview (Barrett et al., 2006), reluctance (Kaminska et al., 2010), social distance between the interviewer and the respondent (Hurtado, 1994), and as explanation for mode differences (Herzog et al., 1988). For example, Barrett et al. (2006) found that 97% of all respondents were rated "as being intellectually capable of participating in the survey, as giving reasonably accurate responses and as understanding the questions being asked" (p. 4028). Despite this near-ceiling level of interviewer-rated ability, those who yielded a poor evaluation were more likely to have item nonresponse on income (but not other) questions, to provide uncodable verbatim responses to open-ended questions, and to vary in socio-demographic and disability characteristics than those with positive evaluations. Methodologically, interviewer evaluations often face much higher correlated variance due to the interviewer than other types of questions (Cleary et al., 1981; O'Muircheartaigh and Campanelli, 1998) requiring the use of multilevel models or accounting for the clustering due to interviewer in analyses.

Table 3 compares each type of paradata that can be captured on four domains: mode, level of aggregation, cost of collection, and ease of collection. Time stamps are available in all modes at low cost and are relatively easy to collect. At the other extreme, behavior codes can only be collected in interviewer-administered surveys, but the collection is difficult, requiring a number of steps from recording and transcribing an interview, identifying a relevant coding scheme, and con-

Table 3. Features of Different Types of Paradata

	Time Stamps	Keystrokes	Mouse Clicks	Behavior Codes	Vocal Characteristics	Interviewer Evaluations
Modes	Face to face Telephone Web	Face to face Telephone Web	Web Telephone	Face to face Telephone	Face to face Telephone	Face to face
Level of aggregation	Action Question Section Survey	Action Question Section Survey	Action Question Section Survey	Utterance Turn Question Section Survey	Word Utterance Question Section Survey	Section Survey
Ease of collection	Easy	Moderate	Moderate	Difficult	Moderate	Easy
Cost of collection	Low	Low	Low	High	Moderate	Low

ducting the coding, and, given the labor involved, can be quite expensive to conduct. Each source of data can provide insights into each stage of the cognitive response process, and thus be useful for understanding measurement error.

4 Differences in Paradata by Modes

The types of paradata that can be captured vary by mode of data collection, driven largely by the software being used for data collection and the people who are interacting with the survey instrument, that is, interviewer or respondent. In this section, we will briefly explore differences in the types of paradata that can be collected by mode of data collection.

4.1 In-person Surveys

Paradata in face-to-face interviews reflects actions by the interviewer and the respondent, with both actors influencing what is captured in the computer (Couper and Kreuter, 2013). As a result, paradata measure both the interviewer and respondent's interactions with each other and with a computer (Couper, 2009). Interviewers have an effect on what is recorded in the computerized instrument. In most face-to-face interviews, the interviewer directly inputs information into the computer. As such, variability across interviewers will lead to variation in what is recorded in the computer. Interviewers affect what respondents report, and respondents affect interviewers' behaviors; some interviewers may probe more and some respondents may be more likely to ask for clarification or definitions than others, potentially affecting response latencies and behavior codes. Variation across respondents will lead to differences in what is recorded by the computer, for example, older respondents generally are slower than younger respondents. Finally, the visual design of the computerized instrument may affect both the interviewer's and respondent's behaviors.

Face-to-face interviews using Computer-Assisted Personal Interviewing (CAPI) software on laptops (traditionally) has allowed researchers to collect a wide variety of time stamps and keystroke data. Today, one of the most commonly used CAPI software programs that also collects detailed paradata in face-to-face interviews is Blaise, in which keystroke files are labeled "audit trails," although other survey software also collect timing and/or keystroke data. Since face-to-face interviews often contain both CAPI and ACASI components, face-to-face survey paradata have been used to examine usability of questionnaires for interviewers (Couper et al., 1997; Couper, 2000; Penne et al., 2002) and potential difficulties encountered by respondents in ACASI (Caspar and Couper, 1997; Bart, 2007; Couper et al., 2009) or interviewer-administered (Couper and Kreuter, 2013) components of a survey. CAPI surveys have also been examined via behavior coding (Cannell et al., 1981), vocal characteristics (Barath and Cannell, 1976; Blair, 1977) and often are a source of interviewer evaluations of the respondent (Davis et al., 2007).

Capture of paradata, especially related to timing, in CAPI instruments has become so routine that novel sets of questions that depend solely on response

timing for interpretation are now being included in survey instruments. For example, the 2008 American National Election Studies included the Implicit Association Test (IAT) , measuring "implicit racism" for black versus white pictures and a replication of the test, but using pictures of Barack Obama and John McCain, the candidates for the 2008 U.S. Presidential Election (DeBell et al., 2010). In the IAT, respondents press keys as quickly as they can in reaction to words or images that appear on a screen. Images or words are selected to represent a target construct (e.g., female or male) and additional words are selected to represent a valenced continuum (e.g., pleasant vs. unpleasant; good vs. bad). The test switches which constructs are paired with which words when displayed on a screen (e.g., "press the "P" key if "female" images or "pleasant" words appear on the screen" vs. "press the "P" key if "female" images or "unpleasant" words appear on the screen") (Lane et al., 2007). Faster response latencies to particular combinations of words and images are interpreted as revealing implicit (i.e., unstated) positive or negative attitudes for one group over another (Wittenbrink et al., 1997).

Because of the physical presence of the respondent during the interview, using the interviewer to collect paradata related to keystrokes and timings is difficult without potentially disrupting rapport between the interviewer and respondent. As such, whatever CAPI computer software is used should capture the relevant information. Blaise routinely captures this information, but other CAPI software programs can capture timing and/or keystroke data this as well (see review of CAPI systems by Shaw et al. (2011». Alternatively, if only timing data are needed, and not direct information about the key entry, the interview can be digitally audio recorded, and timing measured after the interview from the recording itself. These recordings also can be used for behavior coding and analysis of vocal characteristics. For interviewer evaluations, questions are programmed into the CAPI instrument to be answered unobtrusively by the interviewer at the end of the interview or during an ACASI component of the interview. Figure 4 summarizes the steps needed for collecting paradata related to measurement error in a face-to-face or telephone survey.

4.2 Telephone Surveys

The interviewer-respondent-computer interaction for face-to-face surveys also applies to telephone surveys. Notably, in contrast to face-to-face surveys, in telephone surveys, the respondent cannot see the computer. As such, the interviewer's actions using the computer can be more detailed in recording of time and other information about the survey interview.

A wide variety of methods have been employed to capture response latencies in CATI systems, either using the interviewer or the CATI system. When measured by the interviewer, he or she starts a clock as soon as he or she finishes asking the question, and stops the clock when the respondent answers the question, provided a time interval for the "thinking" part of the respondent's response (Bassili and Fletcher, 1991). This timing measure involves interviewer judgment as to when to start the clock when they have finished reading the question and when to stop the clock when the respondent first begins to provide an answer

1. Identify the CAPI or CATI software you will use for your study.
2. Identify the types of paradata your CAPI or CATI software has the built-in capacity to collect-timing, keystroke and/or mouse movement
    a. If the CAPI or CATI software does not have the built-in capacity to collect these data, identify whether ad hoc programs can be added or written for the software. This may require hiring a computer programmer.
    b. If the CAPI or CATI software does collect some or all of these paradata, then identify the necessary programming changes to "turn on" their collection.
    c. Identify where and how the files will be recorded for each respondent.
    d. Develop procedures to export and store the files for each respondent with unique file names.
3. Identify whether you want to collect audio recordings for behavior coding and/or vocal characteristics analysis
    a. Identify whether your CAPI or CATI software has the built-in capacity to obtain record the interview. If not, identify and purchase alternative software to digitally record the interview on the CAPI laptop or tablet or CATI desktop or laptop or obtain separate audio recording device.
    b. Test whether the laptops, tablets or other devices used for the CAPI or CATI data collection have microphones of adequate levels of detection for recording the interview. If not, purchase upgraded compatible microphones.
    c. Identify where and how the files will be recorded for each respondent.
    d. Develop procedures to export and store the files for each respondent with unique file names.
    e. Additional steps for behavior coding
        i. Transcribe the audio recordings to facilitate analysis.
        ii. Review existing coding schemes. Decide if you will code at a question or turn level.
        iii. Develop coding scheme for interviewer and respondent behaviors.
        iv. Hire at least two coders.
        v. Train coders.
        vi. Code interviews, generally coding independently across coders.
        vii. Assess reliability of coding using appropriate statistical methods.
        viii. Identify method for reconciling inconsistencies across coders.
        ix. Reconcile inconsistencies and produce final-behavior coded dataset.
    f. Additional steps for vocal characteristics
        i. Identify software for conducting vocal analysis.
        ii. Split audio files into appropriate analytic units (words, utterances).
        iii. Identify vocal characteristics of interest for your study.
        iv. Use software to conduct appropriate analyses.
        v. Export data into analytic dataset.
4. Identify whether you want to collect interviewer evaluations.
    a. Identify the constructs for which you want interviewer evaluations.
    b. Review existing questionnaires for examples of previously collected interviewer evaluation questions. This includes both question wording and the response options used to collect the information.
    c. As part of the CAPI or CATI instrument, program questions to be answered by the interviewer at the end of relevant sections or the entire survey.
    d. Export evaluations as part of final data instrument. Merge on (masked) interviewer IDs to facilitate analysis accounting for clustering by interviewers.

Figure 4. How to collect paradata for a face to face or telephone interview.

(Bassili, 1996). To account for interviewer variability in these timing measures, voice-activated timers have been used in CATI surveys that start when the respondents make their first utterance (Bassili, 1996). Although the voice-activated timers removed measurement error due to the interviewer, if respondents coughed, requested clarification, or made some other nonverbal linguistic utterances which Bassili called the "hemming and hawing" effect (Bassili and Fletcher, 1991; Bassili, 1993), the timing data are not considered as measuring the time until an answer (i.e., the response latency), and the timing data for that item is often thrown out for that respondent, thereby reducing the analytic sample size substantially. Importantly, correlations between the interviewer judgment and the timer cued by the respondent's voice are between 0.85 and 0.99 in one study (Bassili and Fletcher, 1991), and somewhat lower-between 0.73 and 0.74-in a second study (Mulligan et al., 2003).

Alternatively, additional "hidden" questions can be added to a survey questionnaire as can "latent" timers that begin measuring "time" as soon as the question appears on the interviewer's screen and stop when the respondent's answer is recorded. For "hidden" questions, interviewers are instructed to press a key (e.g., (1)) when they finish reading a question and then press a key (e.g., (1)) when the respondent begins his/her answer. These "hidden" questions add variables to the dataset recording the time of each of these events, but are not read to the respondent (Johnson, 2004; Grant et al., 2010). "Latent timers," as opposed to the more costly "active" interviewer-generated timers, include time for the interviewer reading the question, the time respondents spend thinking about the answer, questions, clarifications, and rapport behaviors, and the respondent's answer (Mulligan et al., 2003), and are identical to those discussed above for CAPI surveys.

As with CAPI software, timing data can be obtained from digital audio recordings (Draisma and Dijkstra, 2004). Also as with CAPI software, these digital audio recordings can be used for behavior coding and analysis of vocal characteristics as well as for timing data.

Keystroke data can also be recorded in CATI surveys. For example, Edwards et al. (2008) examine backing up and data entry errors using keystroke data in a CATI establishment survey. Additionally, analysis of keystroke files are often recommended as a check against falsification of data by CATI (or CAPI) interviewers (Johnson et al., 2001).

Interestingly, as more and more CATI software systems are becoming integrated with web survey software, recording of keystroke data in addition to timing data is becoming less common. One reason for this is that the web software and CATI software are built using the same platform, and web survey systems require JavaScript to capture CSP such as keystrokes or mouse clocks, a feature which is not "turned on" automatically. CATI researchers interested in collecting keystrokes or mouse clicks in an off-the-shelf CATI system that has migrated to a web-based environment will need to implement one of the various JavaScript-based languages when programming the questionnaire, to the extent that it is possible in a particular CATI interface. Of course, use of CAPI software that collects keystrokes (such as Blaise) in a CATI system will also generate keystroke data in a telephone survey.

Interviewer evaluations in telephone surveys appear at the end of the instrument, likely after the respondent has hung up the telephone. These can be easily programmed as additional questions to the interviewer that are simply not read aloud.

4.3 Web Surveys

Web surveys provide a different avenue for paradata collection compared to either in-person or telephone surveys since the mouse clicks and keystrokes are made by the respondent, not an interviewer (Couper et al., 2009). As such, paradata reflect only the respondent's actions. The respondent is also directly influenced by what is presented on the computer, unlike telephone surveys and unlike many face-to-face surveys that do not have an ACASI component.

In a web survey, recordings of mouse clicks made by the respondent are called "client-side paradata (CSP)" whereas recordings of the time that a webpage is submitted to the server on which the survey is hosted are called "server-side paradata" (Heerwegh, 2003, 2011b), see Callegaro, Chapter 11, for a different classification of web survey paradata). Paradata research from web surveys has been facilitated by useful and free JavaScript code permitting unobtrusive collection of CSP written by a variety of European researchers (Kaczmirek, 2008; Stieger and Reips, 2010; Heerwegh, 2011a). Even if one's commercially purchased web survey software does not automatically collect CSP, researchers can implement and tailor the JavaScript for their own use as long as the software permits. These tools facilitate collecting information on the operating system, web browser, screen resolution, respondent time stamps at the survey and item level, and respondent actions on the webpage including accessing drop-down boxes, clicking on radio buttons, changing answers, among other activities (Heerwegh, 2002; Stieger and Reips, 2010; Heerwegh, 2011b). While most web survey software programs collect server-side paradata, such data are usually limited to the number of times the survey's webpage, a time stamp for the visit and (if desired) the respondent's IP address (Heerwegh, 2002; Bowen et al., 2008). The two studies (of which we are aware) that have compared timing data from CSP to timing from server-side paradata have found correlations well above 0.9 (between 0.944 and 0.997, Kaczmirek (2008); between 0.91 and 0.99, Yan and Tourangeau (2008)). However, as with any measure containing measurement error, it is possible that some attenuation of the relationship between response timing data and an outcome of interest occurs when using server-side paradata compared to CSP in analyses (Kaczmirek, 2008).

The web survey software industry is constantly evolving and developing; for example, a 2002 review of web survey software packages (Crawford, 2002) examined three software programs that no longer exist in their current form. A variety of off-the-shelf software packages are available to researchers and their organizations with paradata-collecting capabilities, either built in or through the addition of JavaScript (see overview of possible features in Kaczmirek, 2008). For a list of computerized data collection software, the Association for Survey Computing (http://www.asc.org.uk/) keeps a list (as of this writing last updated in 2006) of software packages, organized by function (e.g., data collection, data analysis).

1. Identify the web survey software you will use for your study.
2. Identify whether your web survey software collects server-side paradata.
   a. Identify where and how the files will be recorded for each respondent.
   b. Develop procedures to export and store the files for each respondent with unique file names.
3. Identify the types of client-side paradata your web survey software has the built-in capacity to collect-timing, keystroke and/or mouse movement.
   a. If the web software does collect some or all of these paradata, then identify the necessary programming changes to "turn on" their collection.
   b. If the web software does not have the built-in capacity to collect these data, identify whether your web survey software permits programming in JavaScript to be added. Add the relevant JavaScript program to the software. This may require hiring a computer programmer.
   c. Identify where and how the files will be recorded for each respondent.
   d. Develop procedures to export and store the files for each respondent with unique file names.

Figure 5. How to collect paradata for a web survey.

WebSM (http://www.websm.org/) also maintains a list of web survey software that may also include data collection tools for other modes. Figure 5 summarizes the steps involved in collecting paradata for a web survey.

5 Turning Paradata into Datasets

5.1 Paradata as Text Files

One of the challenges of working with paradata is the complicated structure in which it is output. In many instances, as shown in Figures 1 and 2, paradata are output as text files that need to be read into a dataset and converted to something analyzable.

The question for data users is often how to turn these somewhat unintelligible strings into useful data files. In web surveys, users of Heerwegh's JavaScript can use his webpage to convert some files to response latencies and measures of answer changes (see http://tinyurl.com/cm5tybp), along with detailed descriptions of how each action is recorded. Users of the Blaise interviewing system can access tools for processing Blaise audit trails developed by users at the University of Michigan and Westat (http://www.blaise.com/Tools). In lieu of these tools, researchers who want to turn paradata into analyzable data must identify variable names (e.g., F1A and F1B in Figure 2 and v[1], v[2_1] and so on in Figure 3), delimiters (e.g., ", :, and $), and actions (e.g., recording of answers "5" and "3" in Figure 2, action of "clicked," "Window Blur" indicating changing screens, and "form submitted" in Figure 3). Then these can be used to create analytic variables through programming in statistical software or through other programming languages such as AWK or Perl.

Parsing the actions from the time at which the action occurred, the time it took for the action to occur, or the time between actions is a challenging task. For exam-

```
Wed Apr 04 2012 14:00:50 GMT-0500 (Central Daylight Time)£
t_fullDownload=6£t=246:WindowFocus£t=280:StartScroll£t=499:EndScroll
at 1475px£t=2569:v[1]=1£t=4005:v[1]=2£t=1114:target=#v2£t=315:StartScroll
£t=493:EndScroll at 3019px£t=494:v[2 1]=clicked£t=1920:v[2 1]=1
£t=554:v[2 2]=clicked£t=1081:v[2 2]=2£t=1727:v[2 2]=3£t=1145:target=#v3
£t=89:StartScroll£t=499:EndScroll at 4426px£t=1841:v[3]=4
£t=848:target=#v4£t=307:StartScroll£t=499:EndScroll at 5894px
£t=3088:v[4]=1£t=1206:target=#v5£t=200:StartScroll£t=513:EndScroll
at 7439px£t=2754:form submitted£
```

Figure 6. Example of output from Heerwegh's CSP project.

ple, in Figure 2, for question 1 (v[1]) the respondent clicked on response option 2 for 965 ms (0.965 s) after the question was displayed, (t=965: v[1]=2), changed his/her answer to option 4 for 2541 ms later (t=2541: v[1]=4) and then finally arrived at the answer of option 1 after another 1566 ms (t=1566: v[1]=1). The question to the analyst is then how to record the data-with an observation for each action resulting in three observations for this question, at the question level, with a variable indicating the total number of actions (3), total number of answer changes (2), or the total time on the question (5072 ms or 5.072 s). Alternatively, the analyst could aggregate over all of the questions in the survey for a single observation per respondent (e.g., adding up all of the t = number values). Given the potentially varying numbers of observations for each respondent and each question, analysts who translate paradata files themselves into data files will need somewhat extensive data management experience using flexible analytic software such as SAS, Stata, R, or syntax-based SPSS rather than transferring the data into Excel or some other spreadsheet program.

Figures 4, 5, and 6 illustrate how to turn a long line of output from Heerwegh's CSP project into data using SAS code. The goal is to create an observation for each action taken in the survey, parse the time for each action from the actions themselves, cumulate time across the entire survey, and count the number of actions for the individual. Figure 6 provides yet another example of paradata output.

As shown in Figure 7, SAS requires the user to indicate a delimiter (dlm="$") and a length of the record (lrecl=600). The length statement indicates that SAS will read in a variable called "action" that will be a character variable of length 60 char-

```
data paradata;
    infile 'CSPexample.txt'dlm='£'dsd lrecl=600;
    length action $60;
    input action $@@;
    time = 1*substr(action, index(action,'=')+1,findc(action,':')-3);
    action1 = substr(action, index(action,':')+1);
    cumtime+time;
    if time=. then delete;
    count+1;
run;
```

Figure 7. SAS code to turn the CSP project paradata file into data.

Table 4. SAS Analytic Dataset from CSP Paradata

action	time	action1	cumtime	count
t_fullDownload=6	6	t_fullDownload=6	6	1
t=246:WindowFocus	246	WindowFocus	252	2
t=280:StartScroll	280	StartScroll	532	3
t=499:EndScroll at 1475px	499	EndScroll at 1475px	1031	4
t=2569:v[1]=1	2569	v[1]=1	3600	5
t=4005:v[1]=2	4005	v[1]=2	7605	6
t=1114:target=#v2	1114	target=#v2	8719	7
t=315:StartScroll	315	StartScroll	9034	8
t=493:EndScroll at 3019px	493	EndScroll at 3019px	9527	9
t=494:v[2_1]=clicked	494	v[2_1]=clicked	10021	10
t=1920:v[2_1]=1	1920	v[2_1]=1	11941	11
t=554:v[2_2]=clicked	554	v[2_2]=clicked	12495	12
t=1081:v[2_2]=2	1081	v[2_2]=2	13576	13
t=1727:v[2_2]=3	1727	v[2_2]=3	15303	14
t=1145:target=#v3	1145	target=#v3	16448	15
t=89:StartScroll	89	StartScroll	16537	16
t=499:EndScroll at 4426px	499	EndScroll at 4426px	17036	17
t=1841:v[3]=4	1841	v[3]=4	18877	18
t=848:target=#v4	848	target=#v4	19725	19
t=307:StartScroll	307	StartScroll	20032	20
t=499:EndScroll at 5894px	499	EndScroll at 5894px	20531	21
t=3088:v[4]=1	3088	v[4]=1	23619	22
t=1206:target=#v5	1206	target=#v5	24825	23
t=200:StartScroll	200	StartScroll	25025	24
t=513:EndScroll at 7439px	513	EndScroll at 7439px	25538	25
t=2754:form submitted	2754	form submitted	28292	26

acters ($60). The input line indicates the variable to read in (action), that it is a character variable ($), and to create a new observation each time the variable "action" is filled (@@). The data step then creates a new variable called "time" using the "action" variable using the substring function (substr). Because the timing data occur after an equals sign (=) and before a colon (:), the index function identifies where to start reading the "action" variable and the findc function identifies where to stop reading the "action" variable; +1 and –3 account for additional characters that need to be ignored in the substring procedure. The substring function yields a character variable; multiplying this variable by I turns the results from a character variable to a numeric variable. Actionl then takes the actions themselves and puts them in a separate variable for future analyses. "Cumtime" creates a cumulative time across the entire dataset-observations without timing data (e.g., the first observation containing date and time) are deleted. Finally, the count variable keeps track of the actions in the order in which they occurred in the survey.

This SAS code in Figure 7 yields the dataset displayed in Table 4. It is non-rectangular, with 26 observations for a single respondent. If a second respondent was added, he/she would have a different number of observations, one for every action he/she took during the survey. The observations start with the time for a full download of the first webpage and end with submitting the final page of the survey.

5.2 Paradata as Sound Files

Behavior codes and vocal characteristics start from recordings, usually digital sound files in today's computer-assisted environment. One of the difficulties in using sound files is parsing the files into small segments of speech. Sequence Viewer (http://www.sequenceviewer.nl/), a program freely distributed for computers with Apple operating systems, allows transcripts and sound files to be linked for purposes of behavior coding, but the individual sound files must be created. Praat, a program used for analyzing vocal characteristics, requires parsing large sound files into smaller segments and then marking individual words or phrases of interest within the smaller sound file for analysis (see description in Groves et al., 2008, pp. 390- 393). Although neither of these tasks is particularly difficult, they are labor intensive, requiring ample numbers of research assistant hours. Even if not linked to sound files, at the bare minimum, behavior codes record from the sound files whether or not an action occurred at a given question (e.g., question read exactly as written) and the actor for the action (e.g., interviewer).

As with paradata from text files, management of data from sound files may be quite involved, especially if multiple behaviors are recorded per question for each actor. If the sequence of the actions is important (i.e., the respondent's request for clarification followed the interviewer's misreading of the question), then the analytic dataset must record action, behavior, and order. To date, there are no examples of analysis of vocal characteristics for multiple items from a single survey, likely due to the data management challenges at hand.

5.3 Paradata as Variables

In yet another form, question-level, section-level, and survey-level response timing variables may be output automatically as part of a dataset, created using "hidden variables" or "timers" added to the questionnaire during data collection. Many public use datasets contain this information already. For example, the public use paradata files for the Consumer Expenditure Survey include section-level timings and the American National Election Studies have included the total interview length in public use datasets since 1964 (Political Behavior Program, 1999). For example, Olson and Peytchev (2007) examine the association between interviewer experience and interview length using the total interview length in the ANES, finding that more experienced interviewers have shorter interviews, on average, than less experienced interviewers. Although this is by far easier than dissecting the text files above, it is also much less rich in detail than the micro-level paradata files described above.

6 Summary

In this chapter, we have described a wide variety of types of paradata, the kinds of paradata available by mode, and some of the challenges involved in turning paradata into analytic variables. These paradata include automatically captured timing data, keystroke data, and mouse click data, and researcher-designed behavior codes, vocal characteristics, and interviewer evaluations.

Given the large amount of data that can be collected, survey researchers' and data collection organizations' decisions about collecting paradata should be driven by a research question or survey management goal. Furthermore, decisions about which types of paradata to collect depend on the mode, budget, time to allocate to analysis, data management skills of the research team, availability of transcribers and/or coders, and storage space. The decision also depends highly on the software package being used by the organization for data collection and the types of paradata it can collect.

This chapter focused on describing the different types of paradata that can be collected by a survey organization. In the next chapter, we explore the analysis of these types of paradata.

Note — This work was supported in part by the National Science Foundation Grant No. SES- 1132015. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

Bachorowski, J.-A. (1999). Vocal Expression and Perception of Emotion. *Current Directions in Psychological Science*, 8(2):53-57.

Bänziger, T. and Scherer, K. R. (2005). The Role of Intonation in Emotional Expressions. *Speech Communication*, 46(3-4):252-267.

Barath, A. and Cannell, C. F. (1976). Effect of Interviewer's Voice Intonation. *Public Opinion Quarterly*, 40:370-373.

Barrett, K., Sloan, M., and Wright, D. (2006). Interviewer Perceptions of Interview Quality. *Proceedings of the ASA, Survey Research Methods Section*, pp. 4026-4033.

Bart, O. (2007). Using Audit Trails to Monitor Respondent Behaviour in an Audio-CASI Questionnaire. *Paper presented at the 11th International Blaise Users Conference (IBUC) 2007.*

Bassili, J. N. (1993). Response Latency Versus Certainty as Indexes of the Strength of Voting Intentions in a Cati Survey. *The Public Opinion Quarterly*, 57(1):54-6l.

Bassili, J. N. (1996). The How and Why of Response Latency Measurement in Telephone Surveys. In Schwarz, N. and Sudman, S., eds., *Answering Questions: Methodology for Determining Cognitive and Communicative Processes in Survey Research*, pp. 319-346. Jossey-Bass Publishers.

Bassili, J. N. and Fletcher, J. F. (1991). Response-Time Measurement in Survey Research a Method for CATI and a New Look at Nonattitudes. *Public Opinion Quarterly,* 55(3): 331-346.

Bassili, J. N. and Scott, B. S. (1996). Response Latency as a Signal to Question Problems in Survey Research. *Public Opinion Quarterly,* 60(3):390-399.

Beatty, P. and Herrmann, D. (2002). To Answer or Not to Answer: Decision Process Related to Survey Item Nonresponse. In Groves, RM., Dillman, D. A., Eltinge, J. L., and Little, R. J. A., eds., *Survey Nonresponse,* pp. 71-85. Wiley and Sons, Inc.

Belli, R. F., Lepkowski, J. M., and Kabeto, M. U. (1999). The Respective Roles of Cognitive Processing Difficulty and Conversational Rapport on the Accuracy of Retrospective Reports of Doctor's Office Visits. In Cynamon, M. L. and Kulka, R. A., eds., *Seventh Conference on Health Survey Research Methods,* Department of Health and Human Services. Centers for Disease Control and Prevention. National Center for Health Statistics, Williamsburg, Virginia.

Belli, R. F., Lee, E. H., Stafford, F. P., and Chou, C.-H. (2004). Calendar and Question-List Survey Methods: Association Between Interviewer Behaviors and Data Quality. *Journal of Official Statistics,* 20(2):185-218.

Benki, J., Broome, J., Conrad, F., Groves, R., and Kreuter, F. (2011). Effects of Speech Rate, Pitch, and Pausing on Survey Participation Decisions. Paper presented at the American Association for Public Opinion Research Annual Meeting, Phoenix, AZ

Biemer, P. P. and Lyberg, L. E. (2003). *Introduction to Survey Quality.* Wiley and Sons, Inc., New York.

Bilgen, I. (2011). Is Less More & More Less? The Effect of Two Types of Interviewer Experience On "Don't Know" Responses in Calendar and Standardized Interviews. Dissertation, University of Nebraska–Lincoln, Lincoln, NE.

Bilgen, I. and Belli, R. F. (2010). Comparison of Verbal Behaviors between Calendar and Standardized Conventional Questionnaires. *Journal of Official Statistics,* 26(3):481-505.

Blair, E. (1977). More on the Effects of Interviewer's Voice Intonation. *Public Opinion Quarterly,* 41(4):544-548.

Bowen, A., Daniel, C., Williams, M., and Baird, G. (2008). Identifying Multiple Submissions in Internet Research: Preserving Data Integrity. *AIDS and Behavior,* 12(6):964-973.

Burrell, T. (2003). First Steps Along the Audit Trail. *Blaise Users Group.*

Cannell, C. F., Miller, P. Y., and Oksenberg, L. (1981). Research on Interviewing Techniques. *Sociological Methodology,* 12:389-437.

Caspar, R. A. and Couper, M. P. (1997). Using Keystroke Files to Assess Respondent Difficulties. *Proceedings of the ASA, Survey Research Methods Section,* pp. 239-244.

Chessa, A. G. and Holleman, B. C. (2007). Answering Attitudinal Questions: Modelling the Response Process Underlying Contrastive Questions. *Applied Cognitive Psychology,* 21(2):203-225.

Cleary, P. D., Mechanic, D., and Weiss, N. (1981). The Effect of Interviewer Characteristics on Responses to a Mental Health Interview. *Journal of Health and Social Behavior,* 22(2):183-193.

Conrad, E. G., Broome, J. S., Benk, J. R., Kreuter, E, Groves, R. M., Vannette, D., and McClain, C. (2013). Interviewer Speech and the Success of Survey Invitations. *Journal of the Royal Statistical Society: Series A (Statistics in Society),* 176(1), 191-210.

Conrad, E. G., Couper, M. P., Tourangeau, R., and Peytchev, A. (2006). Use and Non-use of Clarification Features in Web Surveys. *Journal of Official Statistics,* 22:245-269.

Conrad, E. G., Schober, M. E., and Coiner, T. E (2007). Bringing Features of Human Dialogue to Web Surveys. *Applied Cognitive Psychology, 21(2):165-187.*

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

Couper, M. P. (2000). Usability Evaluation of Computer-Assisted Survey Instruments. *Social Science Computer Review,* 18(4):384-396.

Couper, M. P. (2009). The Role of Paradata in Measuring and Reducing Measurement Error in Surveys. Paper Presented at NCRM Network for Methodological Innovation 2009: The Use of Paradata in UK Social Surveys.

Couper, M. P., Hansen, S. E., and Sadosky, S. A. (1997). Evaluating Interviewer Performance in a CAPI Survey. In Lyberg, L., Biemer, P., Collins, M., DeLeeuw, E., Dippo, C., Schwarz, N., and Trewin, D., eds., *Survey Measurement and Process Quality,* pp. 267-285. Wiley and Sons, Inc., New York.

Couper, M. P. and Kreuter, E (2013). Using Paradata to Explore Item-level Response Times in Surveys. *Journal of the Royal Statistical Society, Series A,* 176(1):271- 286.

Couper, M. P. and Lyberg, L. (2005). The Use of Paradata in Survey Research. In *Proceedings of the 55th Session of the International Statistical Institute,* Sydney, Australia.

Couper, M. P., Tourangeau, R., and Marvin, T. (2009). Taking the Audio Out of Audio-CASI. *Public Opinion Quarterly,* 73(2):281-303.

Crawford, S. C. (2002). Evaluation of Web Survey Data Collection Systems. *Field Methods,* 14(3):307-321.

Crawford, S. D., Couper, M. P., and Lamias, M. J. (2001). Web Surveys. *Social Science Computer Review* 19:146-162.

Davis, J. A., Smith, T., and Marsden, P. (2007). General Social Surveys, 1972-2006: Cumulative Codebook. Technical Report 18, National Opinion Research Center, University of Chicago; online at http://sodapop.pop.psu.edu/codebooks/gss/descriptioncitatioin.pdf

DeBell, M., Krosnick, J. A., and Lupia, A. (2010). Methodology Report and User's Guide for the 2008/2009 ANES Panel Study. Technical report, Stanford University and the University of Michigan.

Dillman, D. A., Smyth, J. D., and Christian, L. M. (2009). *Internet, Mail and Mixed-Mode Surveys: The Tailored Design Method.* 3rd edition. Wiley and Sons, Inc., Hoboken, NJ.

Draisma, S. and Dijkstra, W. (2004). Response Latency and (Para)Linguistic Expressions as Indicators of Response Error. In Presser, S., Rothgeb, J. M., Couper, M. P., Lessler, J. T., Martin, E., Martin, J., and Singer, E., eds., *Methods for Testing and Evaluating Survey Questionnaires,* pp. 131-147. Wiley and Sons, Inc.

Dykema, J., Lepkowski, J. M., and Blixt, S. (1997). The Effect of Interviewer and Respondent Behavior on Data Quality: Analysis of Interaction Coding in a Validation Study. In Lyberg, L., Biemer, P., Collins, M., de Leeuw, E. D., Dippo, C., Schwarz, N., and Trewin, D., eds., *Survey Measurement and Process Quality,* pp. 287-310. Wiley and Sons, Inc., New York.

Edwards, B., Schneider, S., and Brick, P. D. (2008). Visual Elements of Questionnaire Design: Experiments with a CATI Establishment Survey. In Lepkowski, J. M., Tucker, C., Brick, J. M., d. Leeuw, E. D., Japec, L., Lavrakas, P. J., Link, M. W., and Sangster, R. L. eds., *Advances in Telephone Survey Methodology,* pp. 276-296. John Wiley and Sons, Inc., New York.

Ehlen, P., Schober, M. E, and Conrad, E. G. (2007). Modeling Speech Disfluency to Predict Conceptual Misalignment in Speech Survey Interfaces. *Discourse Processes,* 44(3):245- 265.

Fazio, R. H. (1990). A Practical Guide to the Use of Response Latency in Social Psychological Research. In Hendrick, C. and Clark, M. S., eds., *Review of Personality and Social Psychology, Research Methods in Personality and Social Psychology,* volume 11, pp. 74-97. Sage Publications.

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

Fleiss, J. L., Levin, B., and Paik, M. C. (2004). The Measurement of Interrater Agreement, *Statistical Methods for Rates and Proportions,* pp. 598-626. John Wiley and Sons, Inc., New York

Fowler, E. J. (2011). Coding the Behavior of Interviewers and Respondents to Evaluate Survey Questions. In Madans, J., Miller, K., Maitland, A., and Willis, G., eds., *Question Evaluation Methods: Contributing to the Science of Data Quality,* pp. 7-21. Wiley and Sons, Inc.

Fuller, W. (1987). *Measurement Error Models.* Wiley and Sons, Inc.

Galesic, M., Tourangeau, R., Couper, M. P., and Conrad, E. G. (2008). Eye-Tracking Data: New Insights on Response Order Effects and Other Cognitive Shortcuts in Survey Responding. *Public Opinion Quarterly,* 72(5):892-913.

Galesic, M. and Yan, T. (2011). Use of Eye Tracking for Studying Survey Response Processes. In Das, M., Ester, P., and Kaczmirek, L., eds. *Social and Behavioral Research and the Internet: Advances in Applied Methods and Research Strategies,* pp. 349-370. Routledge, New York.

Garbarski, D., Schaeffer, N. C., and Dykema, J. (2011). Are Interactional Behaviors Exhibited When the Self-reported Health Question Is Asked Associated with Health Status? *Social Science Research,* 40(4):1025-1036.

Goudy, W. J. and Potter, H. R. (1975). Interview Rapport: Demise of a Concept. *Public Opinion Quarterly,* 39(4):529-543.

Grant, J. T., Mockabee, S. T., and Monson, J. Q. (2010). Campaign Effects on the Accessibility of Party Identification. *Political Research Quarterly,* 63(4):811-821.

Groves, R. M., O'Hare, B. C., Gould-Smith, D., Benki, J., and Maher, P. (2008). Telephone Interviewer Voice Characteristics and the Survey Participation Decision. In Lepkowski, J., Tucker, C., Brick, J., De Leeuw, E., Japec, L., and Lavrakas, P., eds., *Advances in Telephone Survey Methodology,* pp. 385-400. Wiley and Sons, Inc., New York.

Groves, R. M. (1989). *Survey Errors and Survey Costs.* Wiley and Sons, Inc., New York.

Gutierrez, Christina, Wells, Tom, Rao, Kumar, and Kurzynski, David (2011). Catch Them When You Can: Speeders and Their Role in Online Data Quality. In *Midwest Association for Public Opinion Research Annual Conference.* Chicago, IL.

Hansen, S. E. and Marvin, T. (2001). Reporting on Item Times and Keystrokes from Blaise Audit Trails. Paper presented at the 7th International Blaise Users Conference, Washington, DC, September 12-14, 2001.

Healey, B. (2007). Drop Downs and Scroll Mice: The Effect of Response Option Format and Input Mechanism Employed on Data Quality in Web Surveys. *Social Science Computer Review,* 25(1):111-128.

Heerwegh, D. (2002). Describing Response Behavior in Websurveys Using Client Side Paradata. Paper presented at the International Workshop on Web Surveys held at ZUMA, Mannheim, Germany, October 25, 2002.

Heerwegh, D. (2003). Explaining Response Latencies and Changing Answers Using ClientSide Paradata from a Web Survey. *Social Science Computer Review,* 21(3):360-373.

Heerwegh, D. (2011a). Internet Survey Paradata. In Das, M., Ester, P., and Kaczmirek, L., eds., *Social and Behavioral Research and the Internet. Advances in Applied Methods and Research Strategies,* pp. 325-348. Taylor and Francis.

Heerwegh, D. (2011b). The CSP Project Webpage. Technical Report October 20, 2011; online at https://perswww.kuleuven.be/u0034437/publie/esp.htm

Heerwegh, Dirk and Loosveldt, Geert (2002). Web Surveys. *Social Science Computer Review* 20:10-21.

Herzog, A. R., Rodgers, W. L., and Kulka, R. A (1988). Interviewing Older Adults. Mode Comparison Using Data from a Face-to-Face Survey and a Telephone Resurvey. *Public Opinion Quarterly,* 52(1):84-99.

Hurtado, A (1994). Does Similarity Breed Respect: Interviewer Evaluations of Mexican-Descent Respondents in a Bilingual Survey. *Public Opinion Quarterly,* 58(1):77-95.

International Organization for Standardization (2006). International Standard: Market, Opinion and Social Research-Vocabulary and Service Requirements. ISO 20252.

Jans, M. E. (2010). Verbal Paradata and Survey Error: Respondent Speech, Voice, and Question-Answering Behavior Can Predict Income Item Nonresponse. PhD thesis, University of Michigan, Ann Arbor, MI.

Johnson, M. (2004). Timepieces: Components of Survey Question Response Latencies. *Political Psychology,* 25(5):679-702.

Johnson, T. P., Parker, v., and Clements, C. (2001). Detection and Prevention of Data Falsification in Survey Research. *Survey Research: Newsletter from the Survey Research Laboratory,* 32(3):1-2.

Jurges, H. (2005). Interview, Module and Question Length in SHARE. In BorschSupan, A and Jurges, R., eds., *The Survey of Health, Ageing and Retirement in Europe-Methodology,* pp. 82-87. Mannheim Research Institute for the Economics of Aging. Mannheim, Germany. Online at http://www.share-project.org/t3/share/uploads/tx_sharepublications/Methodology_Ch8.pdf

Kaczmirek, L. (2008). *Human-Survey Interaction: Usability and Nonresponse in Online Surveys.* Dissertation, University of Mannheim; https://ub-madoe.bib.uni-mannheim.de/2150/

Kaminska, O., McCutcheon, A. L., and Billiet, J. (2010). Satisficing Among Reluctant Respondents in a Cross-National Context. *Public Opinion Quarterly,* 74(5):956-984.

Kreuter, E and Casas-Cordero, C. (2010). Paradata. *Working Paper Series of the Council for Social and Economic Data (RatSWD),* No. 136.

Krosnick, J. A (2002). The Causes of No-opinion Responses to Attitude Measures in Surveys: They are Rarely What They Appear to Be. In Groves, R. M., Dillman, D. A., Eltinge, J. L., and Little, R. J. A. eds., *Survey Nonresponse,* pp. 87-100. John Wiley and Sons, Inc., New York.

Landis, J. R. and Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. *Biometrics,* 33(1):159-174.

Lane, KA., Banaji, M. R., Nosek, B. A., and Greenwald, A. G. (2007). Understanding and Using the Implicit Association Test: IV: What We Know (So Far) about the Method. In B. Wittenbrink and N. Schwarz (Eds.), *Implicit Measures of Attitudes:* The Guilford Press.

Malhotra, N. (2008). Completion Time and Response Order Effects in Web Surveys. *Public Opinion Quarterly,* 72(5):914-934.

Mangione, T. W., Fowler, E. J., J., and Louis, T. A. (1992). Question Characteristics and Interviewer Effects. *Journal of Official Statistics,* 8:293-307.

Mathiowetz, N. A (1998). Respondent Expressions of Uncertainty: Data Source for Imputation. *Public Opinion Quarterly,* 62(1):47.

Mathiowetz, N. A. and Cannell, C. E (1980). Coding Interviewer Behavior as a Method of Evaluating Performance. In *American Statistical Association Proceedings of the Section on Survey Research Methods,* pp. 525-528.

McClamroch, Kristi J. (2011). "Evaluating the Usability of Personal Digital Assistants to Collect Behavioral Data on Adolescents with Paradata". *Field Methods* 23:219-242.

Mulligan, K, Grant, T., Monson, Q., and Mockabee, S. (2003). Response Latency Methodology for Survey Research: Measurement and Modeling Strategies. *Political Analysis,* 11(3):289-301.

Oksenberg, L., and Cannell, C. E (1988). Effects of Interviewer Vocal Characteristics on Nonresponse. In Groves, R. M., Biemer, P., Lyberg, L., Massey, J. T., Nicholls II, W. L., and Waksberg, J. eds., *Telephone Survey Methodology,* pp. 257-269. John Wiley and Sons, Inc., New York.

Oksenberg, L., Coleman, L., and Cannell, C. E (1986). Interviewers' Voices and Refusal Rates in Telephone Surveys. *Public Opinion Quarterly,* 50(1):97-111.

Olson, K and Bilgen, I. (2011). The Role of Interviewer Experience on Acquiescence. *Public Opinion Quarterly,* 75(1):99-114.

Olson, K and Peytchev, A (2007). Effect of Interviewer Experience on Interview Pace and Interviewer Attitudes. *Public Opinion Quarterly,* 71(2):273-286.

O'Muircheartaigh, C. and Campanelli, P. (1998). The Relative Impact of Interviewer Effects and Sample Design Effects on Survey Precision. *Journal of the Royal Statistical Society, Series A,* 161(1):63-77.

Ostergren, Jason and Youhong Liu (2010). BlaiseIS Paradata. *Blaise Users Group.*

Penne, M. A., Snodgrass, J., and Barker, P. (2002). Analyzing Audit Trails in the National Survey on Drug Use and Health (NSDUH): Means for Maintaining and Improving Data Quality. International Conference on Questionnaire Development, Evaluation, and Testing Methods (QDET), November 14-17, 2002.

Peytchev, A., Couper, M. P., McCabe, S. E., and Crawford, S. D. (2006). Web Survey Design. *Public Opinion Quarterly,* 70(4):596-607.

Political Behavior Program, the Survey Research Center of the Institute of Social Research, U. (1999). American National Election Studies, 1964 Pre-Post Election Study. Technical report, University of Michigan, Center for Political Studies.

Presser, S., and Blair, J. (1994). Survey Pretesting: Do Different Methods Produce Different Results? *Sociological Methodology,* 24:73-104.

Roßmann, Joss (2010). Data Quality in Web Surveys of the German Longitudinal Election Study 2009. In *3rd ECPR Graduate Conference.* Dublin City University.

Schaeffer, N. C. (1991). Conversation with a Purpose—or Conversation? Interaction in the Standardized Interview. In P. Biemer, R. M. Groves, L. Lyberg, N. A. Mathiowetz and S. Sudman, eds., *Measurement Errors in Surveys* (pp. 367-391). New York: John Wiley and Sons, Inc.

Schaeffer, N. C. and Dykema, J. (2011). Response 1 to Fowler's Chapter: Coding the Behavior of Interviewers and Respondents to Evaluate Survey Questions. In Madans, J., Miller, K., Maitland, A., and Willis, G., eds., *Question Evaluation Methods: Contributing to the Science of Data Quality,* pp. 23-39. Wiley and Sons, Inc.

Schaeffer, N. C., Dykema, J., Garbarski, D., and Maynard, D. (2008). Verbal and Paralinguistic Behaviors in Cognitive Assessments in a Survey Interview. Paper presented at the American Association of Public Opinion Research annual meeting.

Schober, M. E and Bloom, J. E. (2004). Discourse Cues That Respondents Have Misunderstood Survey Questions. *Discourse Processes,* 38(3):287-308.

Shaw, A., Nguyen, L., Nischan, U., and Sy, H. (2011). Comparative Assessment of Software Programs for the Development of Computer-Assisted Personal Interview (CAPI) Applications. Technical report, The World Bank Living Standards and Measurement Study.

Smyth, J. D., Dillman, D. A., Christian, L. M., and Stern, M. J. (2006). Comparing Check-all and Forced-choice Question Formats in Web Surveys. *Public Opinion Quarterly,* 70(1): 66-77.

Sperry, S., Edwards, B., Dulaney, R., and Potter, D. E. B. (1998). Evaluating Interviewer Use of CAPI Navigation Features. In Couper, M. P., Baker, R. P., Bethlehem, J., Clark, C. Z. E, Martin, J., Nicholls II, W. L., and O'Reilly, J. M., eds., *Computer Assisted Survey Information Collection,* pp. 351-365. John Wiley and Sons.

Stern, M. J. (2008). The Use of Client-Side Paradata in Analyzing the Effects of Visual Layout on Changing Responses in Web Surveys. *Field Methods,* 20(4):377-398.

Stieger, S. and Reips, U.-D. (2010). What Are Participants Doing While Filling in an Online Questionnaire: A Paradata Collection Tool and an Empirical Study. *Computers in Human Behavior,* 26(6): 1488-1495.

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

Swanson, G. and Brazer, H. (1959). Detroit Area Study, 1959: The Vitality of Supernatural Experience and a Fiscal Research Program. Technical report. Inter-university Consortium for Political and Social Research.

Tourangeau, R., Couper, M. P., and Conrad, E. G. (2004). Spacing, Position, and Order—Interpretive Heuristics for Visual Features of Survey Questions. *Public Opinion Quarterly,* 68(3):368-393.

Tourangeau, R., Rips, L. J., and Rasinski, K. A. (2000). *The Psychology of Survey Response.* Cambridge University Press.

van der Vaart, W., Ongena, Y., Hoogendoorn, A., and Dijkstra, W. (2005). Do Interviewers' Voice Characteristics Influence Cooperation Rates in Telephone Surveys? *International Journal of Public Opinion Research,* 18(4):488-499.

Wittenbrink, B., Judd, C. M., and Park, B. (1997). Evidence for Racial Prejudice at the Implicit Level and Its Relationship With Questionnaire Measures. *Journal of Personality and Social Psychology,* 72(2):262-274.

Yan, T. and Tourangeau, R. (2008). Fast Times and Easy Questions: The Effects of Age, Experience and Question Complexity on Web Survey Response Times. *Applied Cognitive Psychology,* 22(1):51-68.