GUNNISON

Gunnison Consulting Group, Inc.

USABILITY, COMPARABILITY, AND DATA QUALITY ACROSS MODES AND TECHNOLOGIES IN CENSUS DATA COLLECTION

A DISCUSSION OF RELEVANT FINDINGS AND GAPS IN THE LITERATURE

Dr. Fred Conrad
Dr. Mick Couper

University of Michigan

March 31, 2004

PREPARED FOR: **U.S. CENSUS BUREAU**

TABLE OF CONTENTS

USABILITY, COMPARABILITY, AND DATA QUALITY ACROSS MODES AND TECHNOLOGIES IN CENSUS DATA COLLECTION

A DISCUSSION OF RELEVANT FINDINGS AND GAPS IN THE LITERATURE

We thank Suzanne Fratino and Wendy Hicks for helpful comments, ideas, suggestions and materials throughout the preparation of this document.

INTRODUCTION

This document reports the authors' assessments of the five data collection modes/technologies the Census Bureau is developing for the 2010 Census Short form. Because the amount of directly relevant literature is small, these assessments are based on the authors' judgments as well as available literature. This lack of documentation reflects the new and evolving status of these technologies (with the exception of paper questionnaires), and the emphasis on tasks other than survey response in those evaluations that have been conducted to date.

The document is organized into two parts. Part 1 is structured by the tasks required to answer the short form. These are generally identified by the corresponding question on the form, but in a few cases, like "General Navigation," the tasks cut across questions. Within the discussion of each task, our comments are structured by the five modes/technologies (Mail, Internet, Interactive Voice Response (IVR), Computer Assisted Telephone Interviewing (CATI) and Handheld). If there are issues concerning the task that cut across the modes/technologies, these are listed first. Within each discussion of a mode/technology, we identify potential issues and problems, point to relevant literature, and identify gaps in the research literature that could potentially be addressed in studies carried out at the Bureau of the Census. Note that because the relevant literature is often sparse, we make use of mostly (refereed) conference proceedings as well as traditional peer reviewed journal publications. The "Gaps in the Literature" subsection in each task and mode/technology section will serve as the starting point for a follow-up document that describes the design of possible experiments. If we saw no particular concerns for a particular combination of task and mode/technology, then this is stated.

Part 2 consists of general comments about the technologies themselves and their use in Census data collection. These overall technology comments do not fit within discussions of specific tasks.

PART 1. TASKS AND MODES/TECHNOLOGIES

In reviewing the literature, we discuss 13 tasks. The first three are General Navigation, Entering the Census ID, and Instructions. The remaining tasks are associated with a particular question, identified by a question number. (In most cases, two question numbers are listed: one version for the initial respondent, and another for others members of the household.)

1 GENERAL NAVIGATION

1.1 CONCERNS THAT CUT ACROSS MODES

Topic- vs. person-based navigation: Moore and Moyer (1998a, b; see also Loomis, 1999) compared the conventional person-based interview structure to an experimental topic-based structure in the American Community Survey. Experienced CATI interviewers collected the (demographic) data, either asking all questions (topics) for a given person, looping from one person to the next (the person-based approach) or asking about all persons for a given topic, moving from one topic to the next (the topic-based approach). The Topic-based approach took less time on average to administer and interviewers preferred it. However, respondent reactions differed between households with related persons and those with non-related persons. In particular, related households were more likely to prefer the Topic-based approach, finding it relatively easy to provide information about each person at the same time. For example, when all household members share the same last name, the respondent can indicate this once in the topic-based approach but must repeat it for each resident under the person-based approach. In contrast, unrelated households preferred the person-based approach, although they left more items unanswered under this approach than under the topic-based approach.

A key concern with the person-based approach is that respondents might lose track of which person is the current person. For self-administered approaches, this is particularly challenging because it involves reminding the respondents about which person they are answering at any point in the process. This can likely be achieved by filling in the appropriate name.

Topic-based approaches may be most useful in the hands of an experienced and trained user (interviewer) who is highly familiar with the structure of the instrument; in less experienced hands, the approach may be less effective or even counterproductive. It may be difficult for a novice or one-time user to figure out how to navigate by topic, requiring them to split their attention between the navigation and response task. This is an empirical question that can potentially be addressed in the laboratory with a revised user interface, designed to make topic-based navigation as clear and straightforward as possible.

Finally, it could be that instead of requiring users (either interviewers or respondents) to follow one approach or the other, the best approach might be flexible; enabling the user to follow whatever order is most natural to the respondent. On the other hand, there is a concern that novice users could be overwhelmed by this flexibility. A possibly instructive body of literature concerns event history calendars to help respondents recall major events over the respondent's life course (e.g. Belli, 1998; Belli, Shay & Stafford, 2000). While the response task is quite different than what is required by the Census short form, one of the reasons event history calendars seem to have been successful is that they allow interviewers to capture respondents' memories in whatever order respondent recall them. Belli (personal communication, 2004) is currently experimenting with a web-based implementation for one time users; we believe it will be valuable to monitor the results. Similar work has been done in the Netherlands (Hoogendoorn, & Sikkel, 2002), but on a very limited scale and with a panel of experienced web respondents.

GAPS IN THE RESEARCH LITERATURE

Little is known about the topic- versus person-based distinction with self-administration. These issues are ripe for exploration in the laboratory. One could test versions of paper forms and user interfaces designed to promote (1) person-based, (2) topic-based and (3) flexible reporting. Respondents could be recruited with household characteristics that vary, in particular whether or not the household members are related, or they could answer on the basis of a fictional description of a household in which household members are either related or not. Possible measures might include completion time, missing data, response accuracy, reliability of responses (measured with reinterview techniques) and satisfaction.

1.2 MAIL

COMMENTS AND ISSUES

This mode is person-based by design. In principle, respondents are free to report in whatever order is most natural, but the design discourages this. One concern is that related respondents may find this interferes with their preferred order of responding.

1.3 INTERNET

COMMENTS AND ISSUES

An important design decision is how (and whether) to promote comparability with the paper questionnaire. This might be achieved by creating web pages that are facsimiles of the paper mail questionnaire. For example, one influential text (Dillman, 2000) has advocated designing web questionnaires so that they emulate their paper precursors: "Present each question in a conventional format similar to that normally used on paper self-administered questionnaires" (p. 379). Dillman seems to be recommending that designers try to recreate paper forms because the design principles for this format are well established. However, the web is only superficially similar to paper. While both are visual media, there is far more to the web. For example, it is possible to embed sound, video and animation in a web page and far easier to use images and color on the web than on paper. Plus, the two differ in readability: what is readable on paper may not be on a computer screen and vice versa. The motor actions required to respond are fundamentally different: circling option and writing answers with a pencil share little with selecting menu options, radio buttons or check boxes with a mouse or typing at the keyboard. Beyond this, web respondents may behave differently and have different expectations than those responding to a paper form sent by mail. For example, web users are generally impatient. Hert & Marchionini (1996) observed that many visits to web sites seem to terminate after one page presumably because if users do not immediately find what they are looking for, they leave. In sum, when it comes to designing navigation capability, it is probably a mistake to assume a web version of a paper questionnaire should be designed to look like that questionnaire and have similar functionality. Clearly the relevant empirical studies need to be conducted.

If paper respondents fill out the form in orders beyond a strict person-based sequence, it may be that the comparable web page design makes this sort of flexibility explicit. Of course, even if these respondents behave as the form designers intended and use only the person-based approach, it is quite possible web respondents will demand more flexibility. One can imagine flexible navigation between and within household members. Under the first approach (between household members), a respondent would select a navigational sequence, say topic-based, and then the questionnaire would advance question by question, probably one question per page, with fields for each person on each page. Under this approach, all topics would be displayed on a page (or as many as needed) for one person at a time. Under the second approach (within household members), the respondent is able to navigate along one dimension (e.g. topic) and switch to another at will (e.g. person), or move to a particular question/person combination, i.e. a particular cell in the underlying matrix. The web pages created under these approaches would not replicate the paper form visually, but could potentially support the same navigational preferences that the paper form supports.

The Internet version of the 2000 Census was a single, long scrollable form, with a single submit button at the end of all person-level information. In some respects this replicates the paper form in that there is no navigational capability built into the form beyond the scrolling capability that comes with HTML. There were no edits or routing, increasing the likelihood of missed items, and denying the respondent the ability to save work partway. The navigation in the 2003 Test user interface is far more structured and interactive, displaying one question and collecting one response per page. Navigation was restricted to the person-based approach, although the user can move back to persons for whom data are already collected in any order they choose. The user interface for the 2004 Overseas Enumeration Test offers users considerable backward navigational flexibility. For data already collected, users can select persons (as in the 2003 test interface), questions or question/person combinations in any order they choose.

RELEVANT LITERATURE

Norman (1991) reports that people perform more efficiently, better understand their task, and are more satisfied with broad, shallow menu structures than deep, narrow structures. This could suggest that topic-based is preferable to person-based navigation because the former involves traversing the structure horizontally before drilling down, i.e. involves breadth-first navigation. However this work is quite dated in that it was conducted on character-based as opposed to graphical user interfaces, and did not involve web page navigation at all.

GAPS IN THE RESEARCH LITERATURE

More recently, guidelines have been proposed for designing web page (e.g. Nielsen, 2000) and web site (Rosenthal and Morville, 1998) navigation. However, form-based data collection involves a type of navigation that falls somewhere in between the two.

A laboratory study that would help inform both for the design of the 2010 Census short form and web-based data collection generally, would compare and extend the approaches represented by the 2000, 2003 and 2004 Internet test designs. The extensions would allow flexible forward navigation. Respondents/users would be asked to fill out the three versions of the form based on scenarios that are designed to tax the navigational capability. After completing each form, respondents would rate their experience with the form, and after using all three designs, would rank them along dimensions such as control, flexibility, ease of use, speed of completion, and satisfaction.

In addition to laboratory data, it would be useful to record the navigational choices of test users in the 2005 National Tests. (This would probably have to be collected by client-side software, i.e. that resides on the user's computer, to avoid a server transaction with every user action, and uploaded when the completed form is submitted.) If, for example, it turns out that users primarily use one of the navigational schemes, then it may be appropriate to focus on the development and refinement of just that one.

1.4 IVR

COMMENTS AND ISSUES

The current IVR interface (described in BL07_Census Bureau UI Spec 2.1.doc) does not offer respondents much in the way of navigational choice. While respondents may desire more control over how they navigate through the form – and we do not yet know if they do – it's not clear this will improve their performance or that they will necessarily like what is involved in exercising the choice. Given that the movement is tightly controlled, and there are no visual aids, it may be even more important to provide navigational cues to the respondent throughout the IVR interaction.

RELEVANT LITERATURE

Van Buskirk and LaLomia (1995) evaluated navigation by voice versus keyboard command for standard system-management tasks (like finding files and moving windows) and "hands-busy, eyes-busy" tasks (like

data entry and italicizing text). The authors also tested two kinds of speech recognizers, those requiring discrete speech, in which commands are separated by a brief pause, and continuous speech, with which users can issue several commands without pausing. Although this study did not concern telephone interfaces – users were actually controlling a visual display – it may be relevant for current purposes, as it points out both benefits and costs of spoken navigation.

The authors found that navigation by keyboard took about half as long as navigation by voice[1]. In addition, users preferred the discrete to the continuous recognizer, even though discrete speech is less natural to use. This is because discrete speech reduces the computational demands on the recognizer relative to continuous speech, and as a result, the discrete recognizer performed faster, i.e. recognized the speech in less time, than did the continuous recognizer. Overall, spoken navigation commands were most effective when they were brief and involved small vocabularies, and when the users were more experienced with speech interfaces.

Products like Dragon Naturally Speaking (http://www.scansoft.com/naturallyspeaking/) accurately recognize a broad range of terse spoken commands[2], but they are not speaker independent. That is, each user must train the system on his or her voice and pronunciation for about five minutes. This is a small cost for the recurring workplace use in which these systems are typically operated, but it is impractical for one-time Census respondents.

GAPS IN THE LITERATURE

The tradeoff that needs to be evaluated concerns people's desire for choice on the one hand, and the extra effort to take advantage of that choice on the other. People in multi-person households may want the choice of navigating by person or topic, and all users may want the ability to backtrack and change answers. However, once they discover that explicitly controlling the navigation involves extra effort and time, they may be willing to live with less choice. This is of particular concern, because the IVR interface to the Census short form is something users will interact with just one time. For novice users, the investment that is required to master complex navigation for one interaction is unlikely to seem worthwhile. Of course, for users who have experience talking to computers, these costs may be minimized. And for users who desire the freedom to interact by speaking while using their hands and eyes for other activities, the demands of navigating by speaking might be worthwhile.

One can imagine using Wizard of Oz techniques (e.g. Dahlbäck, Jönsson, & Ahrenberg, 1993) to simulate speech technologies in order to evaluate users' preferences independently of the limits of today's technologies. Under this approach, the user would believe he or she is interacting with a computer through an IVR interface, but in fact, a human (wizard) controls the "system" responses. This approach would make it possible, for example, to simulate a dialogue system that recognizes continuous spoken commands as quickly as is humanly possible, thus separating usability of spoken navigation from current computational and technical obstacles. One outcome could be the decision to continue to develop navigation for the IVR for the short form.

A piece of information critical to the development of IVR is the level of demand for a speech interface. This could be assessed initially with a customer survey administered to those using paper and the web. The description of the capability will be crucial. For example, while 4% of respondents chose IVR in 2003, this number might have been lower (or possibly higher) if the description had made it clear that they would be

[1] Of course, even if the advantage for keyboard navigation were to extend to the Census context, it must first be established that speed of navigation is the primary determinant of user satisfaction before making design decisions. It could be that users desire the ability to navigate by speaking to a computer, and do not expect this to be as fast as navigating by keyboard.

[2] These products are also used for dictation tasks where they must recognize continuous speech. The anecdotal evidence is that, for these tasks, quality is highly variable across users.

interacting with an automated system and not a person. If there seems to be interest in using an automated system, the community of potential users can be more narrowly targeted in subsequent surveys to assess the exact circumstances under which this option would be appealing. This could help design an IVR interface that meets the needs of the respondents who want to use it, rather than hypothetical respondents or all respondents.

COMMENTS AND ISSUES

We are uncertain what navigational options are likely to be supported in the CATI application/instrument, but the work by Moore and colleagues discussed above is especially pertinent because it was carried out with CATI (in the ACS). That work implies that interviewers should be able to adapt the approach to whatever best fits respondents' circumstances (e.g. related versus unrelated household) and preferences.

GAPS IN THE RESEARCH LITERATURE

An issue that warrants investigation in the laboratory concerns how navigational choices are communicated to respondents, if they are communicated at all. It could be that respondents will not understand the choice when described over the telephone, because without visual reinforcement the concept may be too abstract. Alternatively, the idea of reporting in different orders may be quite understandable to most respondents. Finally, it may be that the choice is best left to the interviewer, based on answers to the household roster question (i.e. related or non-related household members) and the respondent's spontaneous order of reporting. This can be assessed by varying who (interviewer or respondent) is given control over navigation, and comparing time to complete the task, accuracy (the "truth" would be contained in vignettes given to respondents), and respondent satisfaction.

1.5 HANDHELD

COMMENTS AND ISSUES

The very small screen size is the key design constraint for Personal Digital Assistant (PDA) use. Because navigation typically involves some sort of screen object or widget to control what is displayed (e.g. scroll bars, arrows, "next" and "previous" buttons, etc.), the challenge is how to implement these without sacrificing too much screen "real estate." If interviewers are to have any control over navigation, which would presumably involve some kind of widget, this could come at the expense of visible content. Some questions, e.g. the race item, involve relatively long lists that could be obscured easily by navigational controls.

RELEVANT LITERATURE

An inventive approach formulated by Kamba, Elson, Harpold, Stamper, Sukaviriya (1996) is the use of translucent widgets, which give the appearance of being layered over content that mostly shows through the widgets. In this way, controls (including those for navigation) are visible without obscuring content. Kamba et al. (1996) instructed their users to navigate to particular newspaper stories (the content) and select text from within those stories. Users could "pass through" the widget and reach the underlying text by holding down the mouse button[3] for a relatively long time (e.g., .5 second). If they intended to click the widget, they were to release the button more quickly. The authors also reversed the position of widgets and text, so that text was on top and translucent and widgets were underneath and opaque. Also, they varied the interval after which a down action was interpreted as a pass-through instruction. Both of these factors affected performance and satisfaction, but the main finding for current purposes was that users quickly learned how to operate the pass-through feature, and could perform their tasks quite adequately. Something similar could be adapted to the non-response follow-up/re-interview task for which the handhelds are used, so that interviewers might navigate while viewing as much content as possible.

[3] Users interacted with the PDA-like hardware using a mouse rather than a stylus. The authors suggest that the results transfer directly to stylus input.

2 ENTERING CENSUS ID

2.1 CONCERNS THAT CUT ACROSS MODES

For purposes of confidentiality, Census suggests having an ID for Internet that utilizes as short a string as possible while still uniquely identifying the respondent. However, it also should not be easily constructed in order to prevent against fraudulent returns.

For this task, an advantage of the Internet over IVR is the ability to present an image of the paper form in a web browser. This can help the respondent locate the Census ID or any other information that might need to be transcribed from the paper form to the web form.

2.2 INTERNET

COMMENTS AND ISSUES/ GAPS IN THE LITERATURE

While apparently no information was captured for invalid responses in 2000 and 2003, it would be useful to know more about the kind of invalid numbers respondents enter in future tests. For example, do respondents enter the wrong numbers entirely, transpose digits, or do something else? Usability tests using the 2003 Internet instrument suggest positive results by collecting the Census ID parsed into three short strings of digits, rather than one long string. Lab participants using the 2003 instrument did not have difficulty entering the Census ID number. However, when respondents make errors at this point, it would be useful to know something about the form of their errors.

The 2000 test results indicated that breakoffs/errors were high at this point. A lab study could help identify what types of errors users were committing with what frequency. This could inform the possible redesign of this item.

2.3 IVR

COMMENTS AND ISSUES

While spoken digits are among the best recognized parts of speech (Davis, Biddulph, & Balashek, 1952), long strings of digits, like the Census ID, which are spoken continuously, are likely to lead to recognition errors. One approach to minimize errors would be to require respondents to enter (speak) digits in three sets (5 digits, 5 digits, 4 digits), as was used in the 2003 test. The prompt to find this on the form should mention that there are three sets of digits, and the ID should be printed in this format on the form. If the system misinterprets one or more of the digits spoken by the user, the success of the interaction will depend on the system's error recovery or repair capability.

RELEVANT LITERATURE

In spoken communication between people, speakers often make mistakes. Part of speaking involves monitoring what one is saying for certain mistakes, which sometimes leads to immediate corrections. Consider this example from Clark (1996, p. 272): "This is one of the things that {uh} one of the many things …." Repairs of this kind seem likely to foil current speech recognition technology. When a respondent utters a string of numbers to a speech recognition system and repairs the utterance mid-stream, it would seem particularly difficult for the recognizer to accommodate this, because there is little or no semantic information for even a "smart" recognizer to exploit. Imagine the following utterance: "6 3 2 {uh} that's 2 3." A human listener might reason that the speaker wishes to reverse the order of the second and third digits, but it's hard to imagine an IVR system built on current technology that could make such a sophisticated inference. The state of the science is more focused on accurately classifying phonemes than on applying discourse principles to interpret repairs.

We suspect that such repairs are ubiquitous in users' interaction with the IVR system, but it is possible they may not be. We know, for example, that users are less disfluent (fewer "uh's" and "ah's") when speaking to (what they believe to be) a computer than when speaking to a person (Bloom, 2000). This suggests that it is essential to catalogue the various respondent utterances leading to breakdowns with the IVR system, and then focus on those which are most prevalent.

GAPS IN THE LITERATURE

A Wizard-of-Oz technique could help determine what level of sophistication would be required to accept ordinary speakers' presentation of digit strings. Unless the speakers are constrained to discrete digits, it may be that very human-like recognizers are required. By simulating the technology with a human (wizard), the necessary level of recognition ability can be identified without actually building the technology.

2.4 CATI

COMMENTS AND ISSUES

It only makes sense to collect Census ID with Inbound CATI, i.e. when the respondent places the call. In these cases, we anticipate few problems, because this is a human-to-human dialogue.

3 INSTRUCTIONS

3.1 COMMENTS THAT CUT ACROSS MODES

The instructions in the 2004 Census Test are lengthy, and likely to be ignored by the respondent. This is especially true with visual presentation, i.e. the paper and web-based questionnaires. In fact, respondents can be forced to acknowledge the instructions within a web browser, but this is no guarantee they will actually attend to them. Spoken presentation (IVR, CATI, Handheld) is harder for respondents to ignore, but if they wish to ignore it, this is likely to reduce satisfaction and presumably increase break-offs. Interviewers in both CATI and Handheld interactions can terminate the presentation if respondents seem to want them to. Human interviewers are also better able than automated systems to detect potential confusion on the part of the respondent, and offer appropriate instructions. As it stands, the IVR system enables respondents to halt the presentation of instructions, e.g., "Say 'stop' at any time if you don't need to hear these guidelines [on who to include in HH count]." This may actually make the instructions comparable across visual and spoken modes. However, depending on the frequency of such "stop" commands, this may under-exploit an opportunity to push instructions to at least some respondents.

GAPS IN THE RESEARCH LITERATURE

An underlying research question is whether respondents can accurately determine whether they need instructions. Because some instructions currently included on the form are relevant to only a small percentage of the population, a new approach to the design of instructions may be to a include preliminary phase that enables respondents to determine if they require the full set of instructions. Lind, Schober & Conrad (2001) found that by including a small part of a definition in the question, respondents were more likely to request definitions when they needed them than if there was no such rewording.

A related research question concerns the design of interfaces that give respondents choices in what parts of the definition they wish to be exposed to. With web-based presentation, different components of the definition can be hyperlinked so that respondents can click on the relevant component. With a speech interface, such decomposition is more challenging. Some sort of speech menu will likely be required, and the design will concern structure: should it be linear or hierarchical? If the latter, how much embedding is necessary? We stress that these are new ideas requiring substantial study.

Possible Analysis. It would be useful to know how many respondents actually did say 'stop' to the IVR system when it was presenting a definition. The data for this presumably already exist from the 2003 test.

4 H1 COUNT OF PERSONS

4.1 MAIL

COMMENTS AND ISSUES

There is an inconsistency in the design of the mail form for the 2004 Census Test, which could confuse respondents. In particular, respondents might wonder about the significance of the border around the instructions and this question, because subsequent questions are not visually framed in this way.

This question also involves a complex concept, "household," and in recognition of this complexity, provides a definition. The concern is that people may be unlikely to read the definition even when they should. If their domestic circumstances are sufficiently atypical, they are at increased risk of responding inaccurately without having read the definition. However, their reading the definition requires that they (1) recognize they need clarification, and (2) are willing to take on the extra reading and thinking.

RELEVANT LITERATURE

Gerber, Wellens, & Keeley (1996) have shown that the Census concept of "usual residence" is unfamiliar to many respondents, who think instead in terms of the concepts of home or permanent address. Partly to address this discrepancy between these everyday notions and the official concept of residence, the Census questionnaire gives some guidance as to who should be counted (e.g., "INCLUDE … foster children, roomers, or housemates") and who should be left out ("DO NOT INCLUDE … college students living away while attending college"). Unfortunately, the research suggests that people often disregard such definitional information. When the instructions are intuitive, respondents don't need them; when they are counterintuitive, respondents often don't follow them.

GAPS IN THE LITERATURE

Considering that people seem likely to ignore the content of definitions, one possible solution would be to turn these into explicit questions, e.g., "Did you include…?" This could increase the chances that respondents actually attend to the definitions.

4.2 INTERNET

COMMENTS AND ISSUES

The primary issue for the Person Count question with web presentation, as with paper, is the use of the definition for "household." The concern is that people will not take advantage of it. One possibility would be to make the definition clickable, i.e. displayed on demand but otherwise hidden.

RELEVANT LITERATURE

We know that people do not frequently take advantage of clickable definitions on the web (Conrad, Couper, Tourangeau, & Baker, 2003), particularly if it involves more than one click. They are more likely to do so when the information seems informative and useful than when it does not. We also know that when they do request definitions, it increases accuracy (e.g. Conrad & Schober, 1999). Respondents seem not to read definitions on the web that are displayed by default (Tourangeau, Conrad, Arens, Fricker, Lee, & Smith, under review), particularly for ordinary words like "residence." This is of particular concern when definitions are wordy and may seem overwhelming or impenetrable to respondents.

More interactive display of definitions may help. For example, Conrad & Schober (1999) have tested an interface in which users could request definitions by clicking, but if they indicated uncertainty through relatively long periods of inactivity, the system offered a definition to them. Coiner, Schober, Conrad, & Ehlen (2002) refined this approach by setting the inactivity threshold (after which the system offered a definition) on the median response times for different user groups. In particular, older respondents required longer periods of inactivity before they were offered help than were younger respondents, so that the ordinary thinking time of older adults was not mistaken for uncertainty. Overall, enabling the system to offer clarification improved response accuracy relative to experimental situations in which obtaining the definition was left to the respondent's initiative.

4.3 IVR

COMMENTS AND ISSUES

Two issues come to mind with IVR administration/collection of the household data. The first is the spoken counterpart of the issues surrounding delivering textual definitions on the web. Will respondents be able to absorb, let alone tolerate, a long, spoken definition? How likely are they to tell the system to "stop," and how will this affect their understanding?

The second issue concerns the wording and style used to elicit household membership. The wording of the retry sequences has a fairly formal tone (despite designing the system with a friendly speaking voice), and includes the response categories that are presumably intended to constrain the respondent's utterances, improving recognition, but which could convey "irritation" by the system. For example: "I still didn't hear you. On February sixth, two thousand three, were you living or staying at the address to which your census materials were mailed? Say yes or press 1, no or press 2." If the respondent believes the system is irritated, this could promote break-offs. Presumably the break-off data are available for such an analysis contingent on the use of such prompts. In the 2003 Customer Satisfaction Analysis, respondents were only moderately satisfied (70%) with the system's recognition of their answers. It is possible that differently worded prompts could have affected this satisfaction figure, even if the system's objective performance was unchanged.

RELEVANT LITERATURE

We know (Bloom, 1999) that respondents find obligatory spoken definitions (i.e. that they cannot terminate) very unsatisfying. They rate this form of delivery lowest in satisfaction relative to definitions they can request. However, obligatory definitions produce slightly higher accuracy than those requested by paid participants. This kind of tradeoff may be worth exploring in the context of the household roster question. However, short form respondents are not paid as in the Bloom (1999) study, and so are probably even less motivated to complete the questionnaire. They may well hang up when given long, unsolicited definitions they cannot turn off. As suggested above, a tally of how often the "stop" command is used would be extremely informative. The ability to turn off the definitions, even if they were originally requested by respondents, would be consistent with the interface design principle of "clearly marked exits" (Molich & Nielsen, 1990): users should be able to end any interaction when they want to without suffering.

There is an inherent tension between constraining answers to be more recognizable on the one hand ('yes' or 'no' is easier to recognize than one of seven responses) and more natural on the other. A study by Hansen, Novick, & Sutton (1996) is pertinent. Based on their experience with the 2000 Census short form, the authors developed a taxonomy of prompt styles, and for each style indicated the likely consequence for recognition and satisfaction. The styles were defined as regions in a space of ten dimensions like politeness, terseness, whether or not the question is decomposed into a filter plus main question ("partial decision tree"), and whether or not the response options are listed. So, one can imagine a terse, polite prompt that uses a filter question or a verbose, impolite prompt that packs everything into one question. The authors use recognition rate, user satisfaction and acceptability of response (behavior coded) to evaluate different styles in several tests, the largest of which involved over 4,000 callers. The value of this work is that it underscores the

range of choices available to IVR system designers for particular questions, and the range of factors to be considered in making such choices. It does not point to a single style that is best in all circumstances.

A worthwhile laboratory study might involve varying prompt style and examining users' behaviors and preferences. Rather than recommending that a range of styles be implemented for such a test, we would advocate the use of Wizard of Oz techniques.

4.4 CATI

COMMENTS AND ISSUES

An alternative to strictly standardized interviewing would involve giving interviewers discretion in whether to volunteer the definitions. In addition, they could invite respondents to ask for clarification.

RELEVANT LITERATURE

We know that when interviewers believe the respondent might misunderstand without clarification, volunteering the definition improves clarification (Conrad & Schober, 2000; Schober & Conrad, 1997; Schober, Conrad & Fricker, 2004). The definitions should be clickable in the CATI user interface.

GAPS IN THE LITERATURE

Given the relatively unskilled pool of agents for conducting CATI interviews, the use of conversational interviewing would be a strong test of the assertion (e.g. Conrad & Schober, 2000; Schober & Conrad, 1997) that ordinary interviewers are capable of judging respondent confusion and interacting until they reach mutual understanding. It may well be that these methods require more skilled workforces, as were in place where the techniques were developed and tested. This raises the question of what exact skills are required to implement the technique. Clearly there is a kind of interpersonal perceptiveness that allows speakers to infer that listeners are confused. There is also subject matter knowledge. It may be that lack of the former may be compensated for by strength on the latter, or that the former depends on the latter, so that the more one knows about the topic, the more attuned one is to the listener's understanding. This suggests a study in which conversational perceptiveness (assuming it can be assessed) and subject matter knowledge are independently varied, and data quality is (somehow) assessed in all four combinations of these factors.

4.5 HANDHELD

COMMENTS AND ISSUES

Creating the roster for large households can involve considerable text entry. If users find Graffiti hard to learn, then the on-screen (virtual) keyboard might be recommended.

In addition, Graffiti may be less effective at recognizing digits than letters. In some versions of Graffiti, users must enter a numerical recognition mode through a stylus action, and in more recent versions, users simply mark on the right side of the entry region. Depending on what version of Graffiti (or other handwriting recognition software) is available to users, the on-screen numerical keyboard may be recommended.

RELEVANT LITERATURE

In one comparison,[4]Ayan, Karagol-Ayan, Huehnert, & Thakkar (2001) observed an advantage in both time and accuracy for tapping a virtual keyboard over writing Graffiti characters on all four experimental trials. However, as users learned Graffiti, their performance improved to the point that it approached performance with the keyboard by the last trial. The authors suggested the trend could lead to superior performance with Graffiti if more tasks had been completed. In another comparison, this time just with experienced Graffiti

[4] The task involved both alphabetic and numerical characters, but these were not explicitly compared.

users, Giambalvo, Frolov & Norouzi (2001) found no difference in time between the two modes, but a preference for Graffiti.

5 H2 ANY OTHERS

5.1 ISSUES THAT CUT ACROSS MODES/TECHNOLOGIES

COMMENTS AND ISSUES
The issues about respondents' use of definitions – including the infrequency of their use – mentioned in conjunction with H1, apply to this question as well.

5.2 MAIL, INTERNET, IVR, CATI, HANDHELD

We foresee no issues with these modes at this time.

6 H3 TENURE

6.1 IVR

COMMENTS AND ISSUES
The collection of this information might be more effective if split into two parts, e.g., "Do you own your home, rent your home or have some other arrangement?" If own, "Is this free and clear or are you still paying a mortgage?" This will enhance recognition of answers and reduce demands on respondents' working memory to maintain this long question.

RELEVANT LITERATURE
Decomposing the question, as suggested above, would follow the "partial decision tree" style of Hansen, Novick, & Sutton (1996). As they point out, this constrains the possible responses, which should increase recognition, but it also increases the number of questions and thus likelihood respondents will break off. Their point is that such tradeoffs – which are well known to survey methodologists – are intensified by the requirements of speech technology and should be carefully weighed.

Graesser, Kennedy, & Wiemar-Hastings (1998) discuss the limits of working memory in grasping long survey questions (pp. 211-212). As they indicate, working memory is limited both in capacity and time, that is: (1) listeners (and readers to some extent) cannot hold more than a few ideas in mind while the remainder of the question arrives, and (2) whatever they hold, they cannot hold it for long. Graesser et al. (1998) generally suggest keeping questions short and breaking longer questions into two or more shorter strings. Similarly, they underscore the importance of using syntax in which the main clause is completed as close to the beginning as possible (so-called right branching syntax,) because this does not require listeners to hold the beginning of the main clause in memory for long. Other syntax (so-called left-embedded syntax) requires listeners to keep the beginning of the question in mind until the clause that it is part of is completed at the end. Because this is not nearly as serious a problem with written language, it raises the possibility that question wording and other characteristics may need to be adapted to particular technologies/modes in order to achieve consistent interpretation across respondents.

Another implication of the Graesser et al. (1996) essay is that speed of IVR delivery may be very important. This is because, if the speech is too slow, the content that listeners store in working memory when the question begins to arrive may no longer be available (i.e. forgotten) by the time the question is completed.

GAPS IN THE RESEARCH LITERATURE
It would be worth experimenting with various speeds of IVR delivery, particularly for long questions. It could be (paradoxically) that some long questions may not need to be decomposed for spoken presentation if the presentation is fast, though the question then is how fast is understandable. Speed may be adjustable on the basis of respondent characteristics like age, and if necessary, so might decomposition of questions into multiple questions.

6.2 MAIL, INTERNET, CATI, HANDHELD

We foresee no issues with these modes at this time.

7 H4 PHONE NUMBER

7.1 IVR

COMMENTS AND ISSUES
As with the Census ID, we suspect this will be more effective if split into subtasks (e.g., first ask area code, then phone number). However, in the Census 2000, phone numbers were captured successfully using Automatic Number Identification (ANI). We understand that Census staff anticipates including that technology in the 2010 application. Including ANI functionality should reduce the number of callers who have to respond to a question collecting their phone number.

For those respondents for whom ANI does not correctly capture the phone number, system designers should carefully consider the exact style of the prompt. One possibility is to name all three components in a single prompt, or provide three prompts (see Hansen et al., 1996). In the 2003 Census Test, the system prompted with a single prompt, "Please tell me your phone number beginning with the area code." Usability testing suggested this format worked well with respondents, and we are glad that the 2003 data will be examined more closely for indications of difficulty in responding.

GAPS IN THE LITERATURE
An issue that could be informed by laboratory investigation concerns the system prompts about how the respondents should answer. The tradeoff seems to be between requiring respondents to remember all components into which the response is to be partitioned (e.g. area code, prefix, etc.), and then delivering the response in this form versus prompting for each component separately.

7.2 MAIL, INTERNET, CATI, HANDHELD

We foresee no issues with these modes at this time.

8 P1_5, P2_1 NAME

8.1 MAIL

COMMENTS AND ISSUES
The instructions for this item are fairly unclear. What if the person(s) who pay(s) the rent lives somewhere else, e.g. the resident's parents pay the rent? Why not just start with the person who has filled out the questions up to this point?

Many names are more letters in length than there are spaces provided; there should be some provision for this either in the instructions or the design, e.g. provide more space than will ever be needed. These space issues become more important on the Internet and with the Handheld interface because, while it is possible to squeeze characters onto a paper form (with unknown consequences for optical character recognition) it is simply not possible to "squeeze" characters into the fixed fields required in most user interfaces.

GAPS IN THE RESEARCH LITERATURE

It could be that providing enough spaces for the letters of very long names compromises character recognition or other aspects of data capture. This could be assessed in a test of the recognition system.

8.2 INTERNET

COMMENTS AND ISSUES

A concern with web data entry – and to some degree with handwritten entry on the mail form – is how to handle accents and special characters (José, Jürgen, etc.). It may be that these are simply not necessary to identify the respondent, and so should not be accommodated. However, if this is the case then this input limitation should be explicitly communicated to respondents so that they don't waste time trying to enter these characters. Moreover, designers and administrators must be prepared for potential resentment by respondents who feel they are not able to fully identify themselves.

One consideration is whether it makes sense to collect all names at once in a roster format. This would only seem to help if the last name can be repeated without retyping, or for large HHs, to remind respondents of those they may have left out. Under these circumstances, this would be a major advance over paper. However, to really reduce redundancy, designers should make a feature available to respondents that enables them to indicate in a single action that all respondents have the same last name. When respondents have different last names, this approach may be at odds with what respondents would prefer to do. It is a variant of the topic-based versus person-based navigation issue.

8.3 IVR

COMMENTS AND ISSUES

Name recognition is likely to be a major problem for IVR. Despite improvements in natural language recognition, names are not part of typical IVR vocabulary, so may need to be spelled out by respondents. We suspect that this is where the most IVR break-offs occurred[5]. These data should be available and would be worth analyzing to see if this is the case.

The special character problem (e.g. José, Jürgen, etc.) mentioned in connection with web-based entry is also likely to be a problem with IVR. In particular, if these do not affect pronunciation, they will not be recognized by the system. However, if they do affect pronunciation, the system is at risk of misspelling the name. Consider the German vowel combination "äu," which is roughly pronounced "oy." Correctly recognizing this spelling poses a serious challenge to IVR designers.

The IVR system starts with first name, but paper lists last name first. If there is no particular rationale for this procedural difference across mode/technology, then it would make sense to standardize the order of

[5] While it does not definitively implicate the name item, the 2003 Response Mode Analysis paper provides suggestive evidence that the item non-response increased from the name item forward. Prior to the demographic series which includes the name item, the level of item non-response was comparable to what was observed with Internet and CATI data collection, and actually lower than paper. However, beyond this point, item non-response increased substantially, possibly due to the burden of spelling names.

collection names across modes/technologies. If there is a reason for doing this, e.g. the current order somehow improves recognition by limiting the probable sounds of last names after particular first names (an unlikely hypothesis), then that should be explicit. A reason for making the order of names in IVR consistent with the order on paper is that respondents might be reading from the paper form when answering the IVR prompts. However, usability testing following the 2003 National Census Test did not indicate that people referred to the paper form when using the IVR. The extent to which inconsistent orders causes interference is testable.

Research in the automated recognition of spoken proper names continues to make advances, though progress is slow. Much of the early progress concerned automated phone calling in which users speak a name listed in the telephone's address book, which is associated with a phone number, and this number is then dialed. This work is not directly applicable here because the list of names typically has been small (in one test the limit was 50), and the recognition system needs to be trained on the speech of specific speakers. The Census context involves hundreds of thousands of names and millions of speakers. More recently, driven the demand to automate support for telephone directory assistance, some progress has been made recognizing speaker-independent names in large (e.g. 10,000) corpora of names. In the late 1990s, the problem was largely framed as spoken letter recognition, i.e. recognition of spelled names (e.g. Hild & Waibel, 1996), but this is not practical for responding in the Census short form. More recently, greater success has been reported with spoken names (as opposed to letters), and a promising analytic approach has been to use the syllable as the unit of recognition rather than phonemes on the small end, and complete names on the large end (e.g. Sethy, Narayanan & Parthasarthy, 2001). While still in the research and development stage, this kind of approach could potentially be used fruitfully with Census respondents. The underlying software differs substantially between applications (neural nets and finite state grammars), and in the speed with which the recognition occurs. If the spoken names can be stored digitally and recognized off line, it could allow slower, more effective technology to be used without slowing down the data collection. Of course, the pervasive problem with off-line recognition is that once the respondent is no longer available, an unrecognizable name is unlikely to ever be recognized correctly. Similarly, names with unusual spellings (e.g. "Couper") will not be recognized correctly without being spelled.

8.4 CATI, HANDHELD

We foresee no issues with these modes at this time.

9 P2_2 RELATIONSHIP TO PERSON 1

9.1 MAIL

COMMENTS AND ISSUES

The wording of the relationship question seems confusing for two reasons. First is the issue of ambiguity of perspective. The relation may be named from the perspective of person 1 or person 2, and the respondent could misinterpret this. Second, non-relatives (as in the last two categories) are odd to encounter in a list of relatives when the question asks about relatives. A "Not Related" label could separate these. (This seems to have been done in the 2004 Overseas web interface.)

9.2 IVR

COMMENTS AND ISSUES

The relationship list is a long list to listen to, which, as Graesser et al. (1998) point out, is likely to be hard for respondents to keep in mind. When the demand on respondents' working memory involves the maintenance

of response options (as opposed to the words comprising the question stem), recency effects – the tendency to pick more recently heard options (e.g. Krosnick & Alwin, 1987) – are more prevalent[6]. (Actually, recency effects have been observed with as few as two response options (see Sudman, Bradburn & Schwarz, 1996, Ch. 6., but they are even more likely with lists of this length).

The steps taken to clarify the relationship on the basis of 2003 usability testing directly aimed at the perspective problem identified above. The system first determines if each person is related to the reference person before asking a more specific question about how they are related, such as, child. In this way, the perspective is established explicitly.

9.3 INTERNET, CATI, HANDHELD

We foresee no issues with these modes at this time.

10 P1_8, P2_5 ORIGIN

10.1 MAIL

COMMENTS AND ISSUES
Respondents may not know what is meant by "Mexican Am," as shown on the Census 2004 Test form. A full label has been requested for subsequent form versions.

10.2 INTERNET, IVR, CATI, HANDHELD

We foresee no issues with these modes at this time.

11 P1_9, P2_6 RACE

11.1 IVR

COMMENTS AND ISSUES
This question poses much the same set of challenges for IVR as did the relationship question. It is again possible that response accuracy would increase if the question were broken into component parts. For example, the respondent would first be asked if she considers herself to be a single race or of multiple races, then would then be routed accordingly.

On the basis of the 2003 usability tests, the question has been reworded to first ask an initial open-ended question with race categories provided as a fallback. This seems like a promising solution to the working memory and response order effects alluded to above.

GAPS IN THE RESEARCH LITERATURE
Respondents of different self-described racial categories, including multi-racial, could be recruited for a laboratory study and presented with either the decomposed set of questions or the single question. Their

[6] Presumably permitting respondents to interrupt (i.e. to select an option as soon as it is heard) will reduce recency effects, but could introduce its own problems.

speed of responding and satisfaction would be the primary measures, where satisfaction would include their judgment about whether they were able to report their race accurately.

11.2 MAIL, INTERNET, CATI, HANDHELD

We foresee no issues with these modes at this time.

PART 2: GENERAL COMMENTS ABOUT MODES

What follows is a set of comments about modes/technologies that cut across individual response tasks. As it turns out, most of our comments at this level concern IVR.

It would be useful to compare completion times for Internet, IVR and CATI administration, conditionalizing on household size and possibly other criteria, such as household members' race/ethnicity and whether or not they are related. The CATI versus IVR comparison, along with the Internet and IVR comparison, is important because it makes it possible to assess how much of the difference between the Internet and IVR is a mode difference (visual versus spoken) and how much is due to technology (self-administered, computerized). Mode comparisons of other, more subjective data, would be useful. For example, in the response mode and incentives experiments (RMIE), about 35% of those who responded to the Internet usage survey gave reasons such as "paper easier/more convenient/prefer paper" for not doing Internet. A further 24% did not have access to the Internet. Unlike IVR, Internet users are relatively young (see the 2003 Response Mode Analysis Paper). Additionally, we suspect Internet respondents are likely to be more educated.

IVR

The IVR dialogue is scripted to convey animacy (i.e. a simulated human interviewer more than a robotic data collector) and informality: "I'm Laura, and I'll be taking down and submitting your Census information." This is a significant design decision, presumably intended to trigger a social reaction within users. Such a decision requires supporting research, because while the anthropomorphic script may in fact lead users (e.g. listing undocumented household members) to treat the IVR system as if it were human, this might be undesirable. First, users (respondents) might be less likely to report potentially sensitive or illegal information to a system that has humanlike qualities (even as superficial as referring to itself as "Laura") than one which is clearly mechanical. In fact, such a reaction might be triggered without anthropomorphic prompts (Reeves & Nass, 1996). Second, this may raise respondents' expectations about the system's capability to unrealistic levels.

Boyce (1999) evaluated user's reactions to four versions of a spoken dialogue for call routing (crossing "I" versus "not I" and casual versus formal), among a total of 84 subjects. Overall, subjects seemed to prefer the version using "I" (although 80% reported not noticing the use of the personal pronoun) and the more casual dialogue. However, these more casual prompts can be wordier, increasing the length of the interaction. Consider the following IVR prompt: "Great, that finishes the first section, now we'll do the section on names. By the way, at any time during this call, you can say 'Tell me something fun about the Census'." This is sufficiently long that respondents may be reluctant to accept the system's offer and endure even more verbiage. The relevant data are presumably available. One wonders if any respondents ever tried this, and if they did, what their experience was.

In general, the authors have mixed feelings about the use of IVR. On the one hand, it is not clear that this technology necessarily reaches a segment of the population who would not otherwise be included. Actually, it's not clear who this population is or if it exists. On the other hand, even if such a population cannot be identified, the IVR system could be redesigned to reach visually impaired respondents. (It requires vision in its current implementation to read the ID and instructions from the paper form.) Additional reasons for skepticism about the viability of IVR comes from its performance in the 2003 test. In particular, the item non-response rate was higher than for paper, and a large proportion of cases required human intervention, reducing the cost benefits. Yet this disappointing performance may reflect the design constraints placed on the 2003 instrument rather than the inherent limitations of speech recognition as applied to a survey or data collection task.

Novick et al. (1999, p. 165) note that "as ATD (automated telephone dialogue) systems typically involve a diverse population of users, they face problems of recognition accuracy beyond those of trainable, speaker-dependent single-user systems." They categorize various systems according to degree of information input (from user to system) and information output (from system to user). They note that questionnaires and surveys are characterized by high information input and low output, in contrast to, say, credit history reports, with low input and high output. They note, "We regard the most challenging tasks to be those that entail high levels of information input." They explicitly mention the census race question as an example of the difficulty of converting tasks that are relatively easy to perform on paper to a verbal form.

In our opinion, if IVR is to be successful, its design must acknowledge and exploit the fact that speech is inherently different from visually presented information. For example, constraining the wording and flow of the census questions to emulate the paper form was a serious obstacle to IVR success in the 2003 National Census Test: Response Mode Analysis: "This constraint resulted in dialogue that was not optimal for an IVR application" (p. 8).

The various steps subsequently taken to improve IVR usability encourage us. These seem largely driven by the specific characteristics of the mode. For example (1) Asking the gender question before the relationship question makes it possible to shorten the relationship question by making it gender-specific. (2) Asking if all persons in the household have the same last name as the first person for whom the respondent provided data to avoid the tedious and time-consuming task of asking for the last name of each household member when everyone has the same last name.

Yet clearly more needs to be done to reduce redundancies in the system prompts. As pointed out in the 2003 National Census Test: Response Mode Analysis (p. 8) "It is also feasible that awkward dialogue may have frustrated respondents to the extent that they hung-up.…" While this kind of user experience strikes us as highly likely the relevant data surely exist and can be analyzed. We recommend examining the IVR data to look at how many repetitions were necessary, how many recognition failures occurred, how many breakoffs occurred at which point, etc.

A final point about the viability of IVR concerns establishing cost criteria up front. According to the 2003 National Census Test Response Mode Analysis report, across all IVR panels, 17-22 percent of IVR returns were ultimately completed by telephone agents (p. 19). This may be an acceptable rollover rate[7] but without a cost-benefit analysis it is hard to interpret. However, we do know that IVR represents a significant savings when compared to agent costs per minute. Therefore, given telephone is offered as a census response mode, any significant percentage of responses captured by IVR provide a net savings compared to the agent alternative. It might also be possible to identify likely rollover cases and transfer them automatically, routing them to appropriately trained agents and reducing the time required of respondents. Customer satisfaction survey results indicated that 76% of respondents who chose to complete the census short form using the IVR in the 2003 National Census Test (over paper or Internet) indicated being satisfied or very satisfied with the system. Additionally, of those same respondents, 77% said they would use the IVR again in the future, suggesting there may be a specific population that we can anticipate using the IVR system. Ideally, we could identify and reflect that population subsequent cost-benefit analyses.

We endorse continued research and evaluation of IVR data collection – including deployment in 2010 – so that when the technology is mature and the design challenges more successfully met, there is a precedent for its use and a body of relevant organizational knowledge.

[7] These percentages are all system failures. If respondents were given a choice of switching to an agent, figures could well have been higher.

REFERENCES

Ayan, N.F., Karagol-Ayan, B., Huehnert, K. & Thakkar, A. (2001). Which is Faster and More Accurate on a Handheld: Graffiti or Keyboard Tapping? *Shore 2001: Student HCI On-Line Research Experiments* (http://www.otal.umd.edu/SHORE2001/graffiti/index.html).

Belli, R. F (1998). The structure of autobiographical memory and the event history calendar: Potential improvements in the quality of retrospective reports in surveys. *Memory*, 6, 383-406.

Belli, R. F., Shay, W. L. & Stafford, F. P. (2001). Event history calendars and question list surveys. *Public Opinion Quarterly*, 65, 45-74.

Bloom, J.E. (1999). Linguistic markers of respondent uncertainty during computer-administered survey interviews. Unpublished doctoral dissertation, New School University, New York City.

Boyce, S. J. (1999). Spoken natural language dialogue systems: User interface issues for the future. In Gardner-Bonneau, D. (ed.), *Human factors and voice interactive systems*. Boston: Kluwer, pp. 37-61.

Clark, H.H. (1997). *Using Language*. Cambridge, UK: Cambridge University Press.

Coiner, T.F., Schober, M.F., Conrad, F.G. & Ehlen, P. (2002). Improving comprehension of web survey questions by modeling users' age. *Proceedings of the American Statistical Association, Section on Survey Methods Research*. Alexandria, VA: American Statistical Association.

Conrad, F., Couper, M., Tourangeau, R. & Baker, R. (2003). Use and non-use of clarification features in web surveys. Paper presented at 58th Annual Conference of the American Association of Public Opinion Research, Nashville, TN.

Conrad, F.G. & Schober, M.F. (1999). A conversational approach to text-based computer-administered questionnaires. In Proceedings of *the Third International ASC conference*. Chesham, UK: Association for Survey Computing, pp. 91-101.

Dahlbäck, N., Jönsson, A., & Ahrenberg, L. (1993). Wizard of Oz studies-why and how. *Knowledge-Based Systems, 6*, 258-266.

Davis, K. H., Biddulph, R. and Balashek, S. (1952). Automatic recognition of spoken digits. *Journal of the Acoustical Society of America, 24*, 637—642.

Dillman, D. (2000) Mail and internet surveys: The tailored design method. New York: John Wiley & Sons.

Gerber, E. (1994). *The language of residence: Respondent understandings and census rules. Final report of the cognitive study of living situations.* Report for the Center for Survey Methods Research. Suitland, MD: U.S. Bureau of the Census.

Gerber, E., Wellens, T., & Keeley, C. (1996). Who lives here? The use of vignettes in household roster research. *Proceedings of the Section on Survey Research Methods, American Statistical Association* (pp. 962-967). Alexandria, VA: American Statistical Association.

Giambalvo, D., Frolov, A. & Norouzi, N. (2001) A comparison of rafitti vs. the on-screen keyboard for experienced Palm users. *Shore 2001: Student HCI On-Line Research Experiments* (http://www.otal.umd.edu/SHORE2001/palmpilot/index.html)

Graesser, A. C., Kennedy, T., Weimer-Hastings, P. & Ottati, V. (1998). The use of computational cogntiive models to improve questions on surveys and questionnaires. In Sirken, M., Herrmann, D.J., Schechter, S., Schwarz. N., Tanur, J. M., Tourangeau, R. (Eds.) *Cognition and Survey Research.* New York: John Wiley & Sons, Inc., pp. 199-216.

Hansen, B., Novick, D. G. & Sutton, S. (1996). Systematic design of spoken prompts. *Proceedings of the ACM Conference on Human Factors in Computing Systems* (pp.157-164).

Hert, C.A. & Marchionini, G.. (1997). Seeking Statistical Information in Federal Websites: Users, Tasks, Strategies, and Design Recommendations. http://www.ils.unc.edu/~march/blsreport/mainbls.html

Hild, H. & Waibel, A. (1996). Recognition of spelled names over the telephone. *Proceedings of the ICSLP 96*, Philadelphia, PA.

Hoogendoorn, A., and Sikkel, D. (2002). Feedback in Web surveys. Paper presented at the International Conference on Improving Surveys, Copenhagen, August.

Kamba, T., Elson, S., Harpold, T., Stamper, T., Sukaviriya, P. (1996). Using small screen space more efficiently. *Proceedings of the ACM Conference on Human Factors in Computing Systems, Interactive Posters* (http://www.acm.org/sigchi/chi96/ proceedings/papers/Kamba/tk_txt.htm)

Krosnick, J.A. & Alwin, D. F. (1987). An evaluation of a cognitive theory of response order effects in survey measurement. *Public Opinion Quarterly, 51,* 201-219.

Lind, L., Schober, M. F. & Conrad, F.G. (2001). Clarifying question meaning in a web-based survey. In *Proceedings of the American Statistical Association, Section on Survey Methods Research*. Alexandria, VA: American Statistical Association.

Loomis, L. (1999), "Nonresponse to Personal Income Questions in Person-Based and Topic-Based Questionnaire Forms." Paper presented at the International Conference on Survey Nonresponse, Portland, OR, October.

Molich, R. & Nielsen, J. (1990). Improving a human-computer dialogue. *Communications of the ACM, 33,* 338-348.

Moore, J., and Moyer, L. (1998a) "ACS/CATI Person-Based/Topic-Based Field Experiment — Final Report." Washington, D.C.: U.S. Bureau of the Census, unpublished report.

Moore, J. & Moyer, L. (1998b). Questionnaire design effects on interview outcomes. *Proceedings of the American Statistical Association, Survey Research Methods Section*, pp. 851-856. Alexandria, VA: American Statistical Association.

Nielsen, J. (2000). *Designing web usability: The practice of simplicity*. Indianapolis: New Riders Publishing.

Norman, K. (1991). *The Psychology of menu selection: Designing cognitive control at the human/computer interface.* Ablex Publishing Corporation

Novick, David G., Brian Hansen, Stephen Sutton, and Catherine R. Marshall (1999), "Limiting factors of automated telephone dialogues." In Gardner-Bonneau, D. (ed.), *Human factors and voice interactive systems.* Boston: Kluwer, pp. 163-186.

Reeves, B., & Nass, C. (1996). *The media equation: How people treat computers, television, and new media like real people and places*. Cambridge, MA: Cambridge University Press.

Rosenfeld, L, & Morville, P. (1998). *Information architecture for the World Wide Web.* Sebastopol, CA: O'Reilly & Associates, Inc.

Sethy, A., Narayanan, S. & Parthasarthy, S. (2002). A syllable based approach for improved recognition of spoken names. ISCA *Pronunciation Modeling and Lexicon Adaptation, 2002*.

Sudman, S., Bradburn, N. & Schwarz, N. (1996). *Thinking about answers: The application of cognitive processes to survey methodology.* San Francisco: Jossey-Bass.

Tourangeau, R., Conrad, F.G., Ahrens, Fricker, Lee & Smith (under review). Everyday concepts and classification errors: Judgments of disability and residence.

Van Buskirk, R. & LaLomia, M. (1995). A comparison of speech and mouse/keyboard GUI navigation. Proceedings of the ACM Conference on Human Factors in Computing Systems, Interactive Posters (*http://www.acm.org/sigchi/chi95/ Electronic/documnts/intpost/rvb_bdy1.htm*).