PROTOCOL FOR PRETESTING DEMOGRAPHIC SURVEYS AT THE CENSUS BUREAU

I. Introduction

The quality of the data collected in surveys is extremely important. One of the activities with the most potential for increasing survey quality is testing the questionnaire before a survey is fielded. In the past, pretesting of questionnaires has tended to be sporadic and inconsistently conducted. It has tended to focus on operational issues rather than on respondents' understanding of the questions. However, since the late 1970's there has been a renewed interest in issues related to measurement error in surveys and more recently, an emphasis on the testing of questionnaires (Schuman and Presser, 1981; Turner and Martin, 1983; Biemer, et al, 1991). There has been a so-called "cognitive revolution," (Jabine, et al, 1984; Loftus, Feinberg, and Tanur, 1985) and with it came the development of research methods that are focused on learning about the cognitive processes used to respond to survey questions.

In addition to interest in the respondent's role in the survey process, there has also been a renewed realization that survey instruments should consist of questions that the interviewers are willing and able to ask as worded (Fowler, 1991; Fowler and Mangione, 1990). Evaluation of the interviewer's role in the data collection process also has implications for how well survey questions provide data that meet the survey's objectives.

Concentration on the contributions of both respondents and interviewers to the quality of survey data has lead to the increased use of cognitive and other methods in the development and testing of questionnaires. For the most part, these methods have been adapted from other areas.

As these methods are used more frequently to evaluate questionnaires, reports of their usefulness are making their way into the literature (Oksenberg, et al, 1989; Campanelli, et al, 1989; Lessler, et al, 1989; Hubbard, et al, 1990; Abraham, 1989; Esposito, et al, 1991). With more and more attention being paid to them, efforts to compare the methods and evaluate their relative effectiveness have also been undertaken (Presser and Blair, 1993).

This monograph describes these newly-emerging methods of pretesting questionnaires for a Census Bureau audience, and provides a set of guidelines for pretesting demographic surveys within the Census Bureau. It is not a "how-to" manual that gives step-by-step instructions for using the new techniques; rather, it is a manual that describes general approaches that can be used in conducting questionnaire pretesting and discusses issues that need to be taken into consideration in planning pretesting activities for any particular survey. With this report, we hope to achieve two main goals. The first is to increase the extent to which pretesting is conducted and make it a routine part of the services we provide to our customers. All too frequently, no

pretesting is conducted before a questionaire is fielded. The second is to modify the Bureau's pretesting procedures so they are consistent with the state-of-the-art in the survey research field.

As the Census Bureau typically conducts them, pretests consist of having interviewers conduct a relatively small number of interviews with a fairly well-developed version of the questionnaire. The number of interviews conducted ranges from nine to several hundred. The questionnaire being tested has either been used previously or has been developed and/or revised by sponsors and subject-matter specialists. In either case, it is considered close to final. Small "hothouse" tests of nine cases are frequently conducted to provide some amount of field testing of the questionnaire in a minimal amount of time. (Nine is the maximum number of interviews allowed without OMB clearance.) Larger pretests are also conducted as time permits. These larger pretests are time-consuming and expensive because they involve OMB clearance, formal forms design and perhaps printing, the same sampling and interviewer training activities required for the actual survey, and sometimes editing and imputation of the data. The main tool for evaluating the results of the pretest is an interviewer debriefing session in which the interviewers report about problems they had in administering the questionnaire, or that they perceived their respondents had in answering the questions. While this provides useful information,

it is not necessarily the best way to find out about problems with the questionnaire. For example, when interviewers report a problem, we do not know whether it was troublesome for one respondent or for many. Also, experienced interviewers sometimes change the wording of problem questions as a matter of course to make them work, and may not even realize they have done so. Thus, they may not always be accurate reporters of questionnaire problems. Given the amount of time and money that goes into fielding a pretest, additional actions taken to improve the quality of the questionnaire being tested and to evaluate the results of the pretest are both extremely worthwhile and cost-effective. The incremental costs of additional pretesting activities, both prior to the pretest as it is currently conducted and evaluating its results, are relatively small compared to the cost of the pretest itself.

The remainder of the report is divided into five sections. In Section II, we present a more detailed treatment of the objectives and scope of a pretest. In Section III, we discuss the techniques that we would like to see incorporated more frequently into the Census Bureau's pretesting procedures. These include cognitive interviews, focus groups, behavior coding, respondent debriefing, interviewer debriefing, split sample tests, and analysis of item nonresponse rates and response distributions. We do not necessarily suggest that all these methods be incorporated in testing of every questionnaire, but

rather that they be considered more widely and used as appropriate. In Section IV, we present guidelines for structuring a pretest plan. There is no one "right" way to conduct a pretest, but rather a number of alternative scenarios depending of the objectives of the testing activity, the amount of time, and the amount of funds available. In this section, we discuss some of the practical considerations that need to be addressed in developing a pretest plan, including time and cost factors, OMB clearance, study design, other pretest implementation issues, reporting of results, and implementation plans for the main survey. In Section V, we include three case studies that were conducted as demonstrations of the use of these expanded pretesting techniques. Prior to the preparation of this monograph, we conducted pretests of three questionnaires, representing a range of situations as far as the length of the questionnaire, time and available funds are concerned. The questionnaires that formed the basis of the case studies were the Current Population Survey (CPS) Supplement on Child Support and Alimony, the Leisure Activities Supplement to the National Crime Survey, and the CPS Tobacco Use Supplement. While full descriptions of the testing activities are presented in the final section, references to these surveys and their testing activities are made throughout the text. In Section VI, we present a summary and conclusions, assessing the strengths and weaknesses of the various techniques.

II. Objectives and Scope of the Pretest

As mentioned previously, there is no one "right" way to pretest a questionnaire. The term pretesting can cover a wide range of activities, from a single question revision in an already tested questionnaire to the development of a completely new instrument. The design of a particular pretest will depend upon the objectives of that pretest and the available resources, including time, funds, and available staff.

Pretests are often used to achieve one or more of the following objectives:

- indicate the source(s) of measurement error in a question, set of questions, or questionnaire that has been used previously;
- examine effectiveness of revisions to an existing question, set of questions, or questionnaire that has been identified as problematic, either through interviewer debriefings or high rates of item nonresponse;
- examine a new question, set of questions, or questionnaire in response to the need for data on a particular subject;
- indicate effects of alternative question versions, modes of data collection, etc., on the data collected;
- assess the final version of the questionnaire for skip pattern accuracy, overall respondent burden and understanding, context effects, mode effects, etc.

The objectives of a specific pretest need to be decided by the sponsor and the Census Bureau staff involved in the pretesting activity. These may vary depending on whether the

questionnaire is being newly developed or revised after a previous round of testing.

In the latter case, information exists that should be examined for evidence of problems. Careful review of the questionnaire by subject matter and questionnaire experts may reveal sources of confusion or structural problems with the instrument. Analysis of data may show over- or underreporting compared to aggregate statistics, inconsistencies between items on the questionnaire, or high rates of missing data. Other documentation (observer reports, interviewer debriefing reports, etc.) may outline other problems or suggestions for solutions.

Regardless of the stage of development of the questionnaire, priorities need to be established. What questionnaire problems are the most important to address? Will the testing encompass the entire questionnaire or just parts of it? Will one or several solutions to problems be developed and tested?

As will be discussed in Section III, there are several different techniques one can use in pretesting. A pretest is not limited to a single implementation of a particular technique; rather, some of the most effective pretests involve iterative testing, encompassing various techniques. For example, in developing a new set of questions concerning health insurance, one may first wish to use a focus group to gain an understanding of how people think about their health insurance and the vocabulary they use in discussing various aspects of their

insurance. A second step might be to develop a set of questions and, using cognitive interviewing techniques, test these questions on various respondents, with the objective of refining the questions so they are easily understood by respondents and the cognitive effort needed to successfully answer the questions is reduced. Finally, a small field test, in which the questions are administered under the same conditions as the final design of the study, may be beneficial in determining any particular problems related to the mode of administration.

Often, the design of the pretest will be limited by constraints. In continuing surveys, it may be necessary to take into account the preservation of an on-going time series, in which major alterations in a question or series of questions would break the series. For all surveys, there are ever-present constraints of time and money. With respect to the latter two constraints, several of the techniques described in Section III can be completed under tight time and funding restrictions (see for example, Case Study #1). These limitations should not be seen as insurmountable roadblocks to testing questionnaires.

III. Techniques

In this section, we describe the various methodologies that can be used to identify problems in the interviewing process and target the source(s) of the error. We divide the techniques into two major categories--pre-field and field techniques.

We define pre-field techniques as those that are generally used during the preliminary stages of questionnaire development. They include cognitive laboratory interviews and respondent focus groups. These techniques are used during the time questionnaire designers are wrestling with how to operationalize survey concepts, what wording to use for specific questions, and how to keep the respondent task to a minimum. These techniques provide the questionnaire designer with more in-depth knowledge of respondent understanding of survey concepts and question wording. This knowledge can then be used to determine what information is feasible to collect, to better operationalize the concepts of interest, and to simplify and clarify the wording of questions.

We define field techniques as those used for the evaluation of questionnaires being tested under field conditions, in conjunction with a field pretest. These include behavior coding of the interviewer/respondent interactions, interviewer debriefings, respondent debriefings, split sample tests, and item nonresponse and response distribution analysis. All of these methods have strengths that contribute to improving a questionnaire to identify problems beyond those that typically surface in a field pretest (e.g., skip pattern errors, respondent fatigue, distraction, hostility to the survey, or lack of motivation). Rather, these methods are intended to discover

problems that are more central to the quality of the data collected and to provide means for improving data quality.

None of the techniques are particularly costly to implement. Obviously, the greatest cost is that of actually conducting the field test. Once the commitment to conduct a field test is made, the field techniques for questionnaire evaluation can be piggybacked onto the field test with minimal additional costs.

A. Cognitive Interviewing Techniques

Cognitive sciences provide the survey methodologist and practitioner with a set of theories and techniques for understanding and improving the survey process. If we look at the respondent's task, we see that each respondent must:

(1) comprehend the question;
(2) retrieve information;
(3) judge whether the retrieved information provides an appropriate response to the question and/or estimate the frequency or evaluation of requested information; and
(4) formulate a response.

In looking at the respondent's task from a cognitive perspective, we can use the tools of cognitive psychology to help develop better questions. (The above model does not take into account the motivational aspects of the response process.)

Cognitive laboratory techniques have come to mean a set of tools used to study the response process and for identifying the errors that may be introduced during this process. The goals of cognitive laboratory techniques are to understand the thought

processes used to answer survey questions and to use this knowledge to find better ways of constructing, formulating, and asking survey questions (Forsyth and Lessler, 1991). The use of these techniques prior to a field test can determine whether question wording communicates the objective of the question, can quickly identify problems such as redundancy, missing skip instructions, and awkward wording with only a few interviews (in contrast to the sample size in a field test), and provide information on sources of response error that are usually unseen by the interviewer and the survey practitioner. In addition to identifying the errors, cognitive techniques often provide information toward a solution to the problem.

Among the repertoire of cognitive laboratory techniques are the following (Forsyth and Lessler, 1991):

- Concurrent "think aloud" interviews;
- Retrospective "think aloud" interviews;
- Follow-up probes;
- Paraphrasing; and
- Confidence ratings.

Each of these techniques will be described below. A discussion of other cognitive laboratory techniques is included in Appendix A.

Concurrent "think aloud" interviews consist of one-on-one interviews (either self-administered or interviewer-administered) in which the respondents describe their thoughts while answering

the questions. Respondents are instructed before the interview to "think aloud" and the interviewer or observer guides the respondent during the interview by reminding the respondent to "tell me what you are thinking" or "say more about that." Concurrent think aloud interviews are often recorded (audio or video) to permit analysis of the session post interview. Think-aloud sessions are often used to identify difficulties in question comprehension, (mis)perceptions of the response task, identification of the types of recall strategies used by respondents, and reactions to sensitive questions.

Retrospective think aloud and probe questions are less time-consuming than concurrent think aloud techniques. In a retrospective think aloud, the respondent first completes the interview, similar to the conditions under which most respondents would complete the interview task. Following the interview, the respondent and interviewer review the survey responses and the respondents are asked about the process used to generate their answers. Interviewers can use general probes (e.g., "tell me what you were thinking when you....") or very specific probes (e.g., "Why did you report the $142 payment as child support?") to guide the think aloud process. In some cases, the researcher/interviewer may choose to audio or videotape the interview and then review the interview with the respondent while asking probing questions. Think aloud interviews and probing questions are similar techniques. In think aloud interviews, the

interviewer can focus the respondent on either the total interviewing process or on specific topics. For example, in think aloud interviews, the interviewer may ask about the procedures that were used or how the respondent, in general, retrieved the information. Probing questions are used when the researcher wants to focus the respondent on particular aspects of the question-response task. For example, the interviewer may ask the respondents how they chose among response choices, how they interpreted reference periods, or what a particular technical term meant.

All of the techniques described above--concurrent and retrospective think aloud interviews, and detailed probing--are designed to assess all aspects of the response formation process (comprehension, retrieval, judgment, and response). The concurrent think aloud has the advantage of capturing the information at the time it is first available to the respondent; however, concurrent think aloud interviews are both time-consuming and may bias responses to questions later in the interview. Retrospective techniques and post-interview probing, especially when either audio or video taped, provide an unbiased means for capturing the data, while still preserving the opportunity for focusing on general or specific questions concerning the interview. However, the respondent may not be able to recall his/her thought processes when asked about them at the end of the interview rather than after each question.

The remaining techniques--paraphrasing and confidence ratings--differ from the think aloud and probing techniques described above in that they focus on a specific aspect of the response formation process. Paraphrasing is simply asking the respondent to repeat the question in their own words. Paraphrasing permits the researcher to examine whether the respondent understands the question and interprets it in the manner intended. It may also reveal better wordings for questions, if different respondents consistently use the same terminology. In research on the Continuing Survey of Food Intakes by Individuals, respondents were asked to paraphrase a question about whether a food item eaten the previous day had ever been brought into their home. Respondents misinterpreted the question as asking about the general type of food item (e.g., hamburger), rather than the one they had eaten the previous day. Based on this finding, the question was revised.

In confidence ratings, respondents answer the survey questions and then are asked to rate how confident they are with their responses. This technique identifies questions that respondents find difficult to answer (low confidence ratings of the response). Low confidence ratings often arise from either lack of knowledge (especially among proxy respondents) or a difficult recall task. While low confidence ratings may provide an indication of the respondent's perceived level of difficulty, they do not necessarily mean that the respondent's answers are

inaccurate. Similarly, high confidence ratings do not necessarily imply accuracy of response.

Conducting Cognitive Interviews at the Census Bureau

These techniques have been used at the Census Bureau to address measurement issues in a variety of surveys. (As the need arises, the range of techniques may be expanded.) The basic approach has been to integrate concurrent think aloud interviews with follow-up probes, paraphrasing, and confidence ratings. For the rest of this discussion, this combination of methods is referred to as "cognitive interviews."

Cognitive interviews have several advantages in the early development of a questionnaire (and even in the testing of previously developed questions). First, they provide an important means of finding out directly from respondents about their problems with the questionnaire. Second, cognitive interviewing requires a small sample size for diagnosing problems. With as few as fifteen interviews, major problems can surface if respondents repeatedly identify the same questions and concepts as sources of confusion. Third, because the sample sizes are small and contained, iterative pretesting of an instrument is often possible. After one round of interviews is complete, researchers can diagnose problems, revise question wording to solve the problems, and conduct additional interviews to see if the new questions are less problematic.

Cognitive interviews are generally conducted with "convenience" samples of respondents. They are not designed to be representative, but to reflect the detailed thoughts and problems of the few respondents who participate in them. While cognitive interviews are sometimes conducted in various regions of the country, typically they are conducted locally due to cost factors. Due to the lack of geographical representation, some problems that might exist elsewhere, due to regional differences in meaning of words or phrases, may not surface during the cognitive interviews.

Respondent recruitment is an important aspect of the conduct of cognitive interviews. Generally, recruiting is accomplished through advertisements or contacts with local organizations. The specific recruiting method will depend on the topic of the questionnaire being tested. For questionnaires that involve the general population, broad-based advertisements in newspapers or pamphlets can be used. Care should be taken to ensure that all types of respondents are included in the interviews. For example, in the CPS Tobacco Use Study (Case Study #3), respondents were screened for smoking status prior to setting up appointments, so that equal numbers of interviews would be conducted with current smokers, former smokers, and non-smokers. When questionnaires involve rare populations, recruiting efforts should focus on more specialized organizations. In the CPS Alimony and Child Support Supplement (Case Study #1), Parents

Without Partners was contacted as a source for locating single mothers.

Appendix B, originally developed by Judith Lessler of Research Triangle Institute, provides a guide for conducting cognitive interviews. Before interviewing on a particular study begins, the staff who will be conducting the interviews get together as a group to plan the protocol--that is, the probes that will be used. These probes can include paraphrasing requests, confidence ratings, or requests for respondents to define terms or elaborate on how they arrive at their answers. The specific content of the protocol will depend on the interviewers' knowledge of actual or potential questionnaire problems, and the objectives of the testing. Every interview will be conducted a little differently because of individual differences between respondents, but the goal is to collect the same types of information from all respondents. This will enable the interviewers to determine the extent to which similar problems, recall strategies, interpretations, etc., were experienced by respondents.

Conducting cognitive interviews requires that the researcher be familiar both with the content and intent of the questionnaire and with techniques for conducting cognitive interviews. Training to use these types of techniques usually requires one to two days, after which the trainee should be

supervised until he or she feels comfortable using the techniques.

Cognitive interviews are frequently conducted in a laboratory. (CSMR's Response Research Laboratory has facilities for conducting interviews as well as for video and audiotaping the proceedings.) However, while the laboratory setting can be beneficial for recording the interviews, it is not a prerequisite. In some cases, potential respondents to a particular survey may not be located near the laboratory. In other cases, respondents may need records or other materials that are located at home or in the office.

Regardless of the location, the interviews are typically tape-recorded (or videotaped, if the opportunity allows) with the respondent's permission. Audiotapes are necessary references for use in compiling summaries of interviews; videotapes and audiotapes are useful for demonstrating problems with the questionnaire to sponsors or staff involved with the survey.

After each interview is conducted, a summary is prepared. This is not a verbatim transcript, but rather an item-by-item description of the events of the interview--whether the question was read as worded, whether the respondent had any problem understanding the question (and if so, what the problem was), how the respondent interpreted key concepts, what strategies the respondent used to come up with an answer, etc. The summaries

should be coordinated with the protocol, so that the same types of information are available for all respondents.

Cognitive interviews were used in all three of the pretests described in the case studies. Evaluation of the usefulness of the information gained during these sessions are included in the write-ups of these pretests. See Section V.

B. Focus Groups

Focus group interviews bring several people together to discuss selected topics. Generally, groups are constructed to represent subgroups of a survey's target population, and several groups may be necessary to represent different types of people since focus groups are often selected to be relatively homogeneous. Generally, the group sessions last for one to two hours, are audio or video recorded, and are led by one or two moderators who guide the discussion by focusing it on particular topics of interest. Focus groups can be used in a variety of ways to assess the question-answering process. These can include:

- as a means for group administration of a questionnaire (usually self-administered) followed by a discussion of the experience;
- as a means for gathering information before the construction of a questionnaire, ranging from how people think about a topic to their opinions on the sensitivity or difficulty of the questions; and
- as a means for quickly identifying variation and homogeneity of language or interpretation of

questions and response options. Used in this way, focus groups provide quicker access to a larger number of people than may be possible with cognitive interviews.

One of the main advantages of focus groups is the opportunity to observe a large amount of interaction on a topic in a limited period of time. The interaction is of central importance--the focus group provides a means for group interaction to produce information and insights that may be less accessible without the interaction found in a group (Morgan, 1988). Another advantage to group interviewing is that exploratory research requires less preparation than a formal interview. The focus group permits observation of a larger number of participants than may be possible with cognitive interviewing techniques (within the same time frame). Time is also saved in the analysis, since fewer transcripts are generated from the focus group process. However, the focus group, because of the group interaction, does not permit a good test of the "natural" interviewing process, nor does the researcher have as much control over the process as would be true with either cognitive interviewing or interviewer administered questionnaires.

The following questions need to be addressed when planning to conduct focus group research:

What Topics Should be Covered? Unlike other survey situations, the researcher needs to be sensitive to additional ethical issues surrounding a topic being considered for focus

group research. First, since the primary means of collecting the data is the (audio or video) recording of the session, decisions as to data access should be made prior to the session. Second, since by its very nature the focus group requires group discussion, each participant will be sharing his or her ideas not only with the moderator (researcher) but also with the other participants. For that reason, topics should be limited to those which can be discussed in a group setting. Not only will sensitive topics present problems concerning invasion of privacy, they will also serve as barriers to free-flowing discussion.

How Many Groups Should be Interviewed? In part, the number of groups will be driven by the research goals and the number of different subgroups of interest. The more homogeneous the target population, the fewer the number of groups needed. One group, however, is never enough, since without the experience of a least two groups it will be impossible to know if the findings from the one group are generalizable or simply the result of the dynamics among a particular set of participants. One guide as to how many groups are needed is to determine whether additional discussions are producing new ideas. If the moderator can anticipate what will be said in the group, there is no need for further groups. In planning focus group interviews, a target number of groups should be determined, with flexibility built in to add or discontinue as the research findings evolve.

What Size Group Should be Used? Groups that are too small (less than 4 participants) will be both less productive and more costly (given a set total number of desirable participants); too large a group (more than 12 participants) will lead to group management problems. In very small groups, each participant will have a greater burden to comment while large groups may encourage silence on the part of some participants since the group as a whole can carry the discussion. While all groups can be destroyed due to the excessive dominance of one participant, the size of a large group will necessitate increased levels of involvement by the moderator. This is not necessarily a desirable characteristic. Given the need to balance substantive and practical problems, it is recommended that focus groups consist of four to twelve participants.

How Much Time Will be Necessary to Prepare for and Conduct the Focus Groups? The actual focus group interview should be scheduled to last between one and two hours. Recruitment of participants may be quite time-consuming when specialized populations are needed. Also, even though the individual focus groups take only a few hours to conduct, conducting more than one per day or three to four per week is considered difficult for a single moderator. This is due to the intensity required to moderate focus groups, as well as the need to prepare some notes after each group has been conducted.

Should a Moderator Guide be Prepared? It is often helpful, especially when different moderators are conducting the focus groups, to prepare a moderator guide. The structure that a guide imposes on discussions is valuable both in channeling the group interaction and making comparisons across groups during analysis. A good guide should create a natural progression across topics with some overlap between the topics. Guides should be developed initially around a full list of the questions of interest; however, general topics should be introduced to the participants rather than specific questions, to avoid restricting the direction of the discussion. Moderators should commit the outline of the guide to memory to avoid referencing written text during the session--doing so slows down the pace of the group and focuses the participant's attention on the goals of the researcher rather than on the interaction taking place. The guide is meant to literally serve as a guide; the moderator needs to be free to probe more deeply when necessary, skip over areas that have already been covered, and follow completely new topics as they arise.

How Should the Focus Group Begin? The session should begin by introducing the topic in a fairly general fashion. By providing a general introduction, the participants are not restricted in their thinking or discussion about a topic. In addition to the general introduction, the moderator should present some rules for the discussion, such as only one person

speaking at a time, encouraging the participation of all individuals, etc. Moderators should emphasize that the session is being conducted so that the researcher can learn from the participants.

Group discussion typically begins with each participant making an individual, uninterrupted statement, often of an autobiographical nature. This procedure serves as a good icebreaker by not only getting everyone to speak, but by providing both the moderator and the other participants with some basic information about everyone. Opening statements also provide a means for getting everyone to discuss their different experiences and opinions before a group consensus can emerge.

After the opening statements, the movement in the discussion will depend on the level of involvement by the moderator. Low levels of moderator involvement will result in relatively unstructured group discussion that proceeds until the moderator introduces a second topic. A more structured group will also begin with a broad discussion, but usually for the purposes of eliciting specific topics to be followed in detail. Finally, it is important to provide the participants with a clear indication as to when the discussion is ending. In some cases, the moderator may wish to have each participant provide a final summary statement. A sense that the final statement will not be interrupted or challenged may allow a participant to make a

contribution that he or she has been holding back from the open discussion.

Where Should the Focus Group be Held? Choosing a site for the focus group must be a balance between the needs of the participants and the needs of the researcher. The setting must be accessible to the participants and provide a facility in which recording will not be problematic. It is useful to have participants sitting at either a rectangular or circular table to facilitate group discussion and audio taping.

What is the Role of Observers in Focus Groups? Sponsors, researchers, or other observers are welcome to watch the focus group proceedings. They should, however, be located in a separate room behind a one-way mirror so they are not actually part of the group. Participants in the focus group should be informed that they are being observed. Separation of the observers from the participants in the focus group is important because just their presence in the room can have an effect on what the participants are willing to say in the group. In addition, if observers are in the same room, they may be tempted to make comments or ask questions that would affect the discussion either by causing it to go in a different direction or by intimidating the participants. Analysis. The two basic approaches to analyzing focus group data are a qualitative summary and systematic coding via content analysis (Morgan, 1988). The qualitative approach involves review of summaries of

each group, drawing conclusions based on commonalities and differences among the groups, and includes direct quotation of the group discussion in the report of results. The content analysis typically involves setting up a coding scheme to capture specific information about the content of each focus group and using it to code the tapes of the groups. The resulting product is numerical descriptions of the data, which are included in the report of results. In either mode of analysis, it must be recognized that the group is the fundamental unit of analysis, and the analysis will at least begin in a group-by-group progression. A useful strategy is to begin with a detailed examination of one or two groups, developing hypotheses and coding schemes and then applying them to the remainder of the groups.

C. Behavior Coding and Analysis

Another method that is useful to implement during pretesting is behavior coding. Behavior coding, which is the coding of the interchange between the interviewer and the respondent, can provide useful information about the quality of the survey data. A systematic way to capture this information involves coding the interaction that occurs during field interviews on a case-by-case basis to note specific aspects of how the interviewer asked the question and how the respondent reacted. This process, called behavior coding, was first used in surveys to monitor and

evaluate interviewer performance (Cannell, et al., 1975) and subsequently as a tool to evaluate the question-answer process more generally (Mathiowetz and Cannell, 1980; Morton-Williams and Sykes, 1984; and Oksenberg, et al., 1991) and to assess the effects of interviewer behavior on response variance (Groves, et al., 1980). Different variations in the general coding schemes have been developed to address the various uses of the data. For example, knowing that an interviewer needed to probe for an adequate answer is sufficient if the purpose of the coding is to identify problem questions; however, if the purpose of the coding is to assess interviewer quality, the coding scheme would need to reflect the adequacy of the interviewer's probe. Both goals may be important in using behavior coding as part of pretest activities.

Figure 1 contains a simplified version of earlier coding schemes, designed for the purposes of a pretest. The coding scheme focuses on the interviewer's and the respondent's behavior indicative of problems with the question, the response categories, or the respondent's ability to form an adequate response. The small number of codes permit relatively fast coding of taped pretest interviews; however, the ease of the coding scheme comes at the cost of no information concerning the interviewer's behavior apart from how the question was initially read. In contrast, Appendix C provides the coding scheme originally developed by Cannell, et al (1975). This scheme is

quite complex and evaluates the appropriateness of each of the interviewer's behaviors as well as their pace and intonation. In developing a coding scheme to be used in a pretest, the researcher must focus on the codes necessary for the evaluation, without unduly burdening the coders.

FIGURE 1. Interview Behavior Codes (Simplified Version)

Interviewer Question Reading Codes

E (Exact): Interviewer reads the question exactly as printed.

S (Slight Change): Interviewer reads the question changing a minor word that does not alter question's meaning.

M (Major Change): Interviewer changes question such that the meaning is altered. Interviewer does not complete reading the question. Interviewer skips a question that should have been asked.

V (Verification): In lieu of asking the question, the interviewer accurately verifies/repeats relevant information that the respondent had provided earlier.

WV (Wrong interviewer Verification): In lieu of asking the question, the attempts to verify pertinent information, but in doing so, presents the respondent with inaccurate or inappropriate information. Also to be used if the interviewer silently records the answer to a particular question based on information the respondent has already provided (and thus does not verify the information with the respondent).

Respondent Behavior Codes

IN (Interruption reading with with Answer): Respondent interrupts initial question-answer.

CL (Clarification): Respondent asks for or clarification of question or makes a statement indicating uncertainty about question meaning.

AA (Adequate objective. Answer): Respondent gives answer that meets question

QA (Qualified objective, Answer): Respondent gives answer that meets question but is qualified to indicate uncertainty about accuracy.

IA (Inadequate meet question Answer): Respondent gives answer that does not objectives.

DK (Don't Know): Respondent gives a "don't know" or equivalent answer.

RE (Refusal): Respondent refuses to answer the question.

Although relatively simple coding schemes can be used "live", that is, during the conduct of an interview, it is recommended that interviewers tape record pretest interviews to be used in coding. Taping permits a more efficient means for coding (e.g., coder's time is not lost due to waiting for an interview to occur) while also facilitating higher quality codes, since a coder has more than one opportunity to listen to the interaction and record what has occurred.

In addition to using the data generated from the coding of pretest interviews to identify problem questions, as part of the pretesting activities it is useful to debrief the behavior coders. Unlike interviewers, the coders have little vested interest in how a question was read or the situation which resulted in a series of interviewer and respondent behaviors. Having listened to a number of interviews within a relatively short time (without the personal involvement of being the person who administered the questionnaire), the coders can provide useful insight into potential sources of the problems.

Behavior coder training usually requires two to three sessions, each lasting between three and four hours. The sessions are intended to introduce the codes, have coders practice in group sessions, and then individually code interviews and compare the results. Coders need to be familiar with the objective of the survey questions; various studies have used researchers and questionnaire designers, interviewers not

involved in the pretest (but familiar with the pretest questionnaire), or general coding staff for behavior coding. Depending upon the staff used to complete the coding, questionnaire and item-specific objectives will need to be reviewed.

Training should be designed to cover the following:

- description of each of the codes and how they are used;
- group demonstration in which the instructor plays short segments of the tape and initiates group discussion of the behavior observed and how to code the behavior;
- group coding of short segments of an interview, followed by group discussion;
- group coding of an entire interview, followed by group discussion;
- individual coding of two or three interviews (this is often done as a homework assignment, requiring the instructor to have prepared several copies of taped interviews); and
- final discussion of individually coded interviews.

Depending upon the consistency among the coders after having coded several interviews individually, the instructor may need to conduct one or more additional group interviews to clarify the use of particular codes.

Appendix D presents a typical coding sheet for behavior coding. This coding sheet was used in conjunction with the codes presented in Figure 1. Coders can simply check the boxes to indicate whether a particular behavior occurred. This coding sheet, and the subsequent analysis, did not provide information

on the sequence in which behaviors occurred, although we usually assume that the question reading behavior was the sequence initiator. Also note that the coder has room to indicate the nature of the problem in greater detail. These notes are invaluable in helping the survey practitioner determine means for improving the question. For example, if the codes indicate that the interviewer read the question making major modification to the question wording, the researcher knows only that there is a problem. If coders have noted that the nature of the problem is that interviewers fail to read all of the response options to the respondent, the researcher can attempt to remedy the situation, through instructions on the questionnaire or in the detailed interviewer instructions.

For any study it is important to measure the consistency among coders, both at the beginning of the coding process and (if it is a long study) throughout the coding, and at the end of the study. Kappa statistics provide a means for assessing the reliability among coders (see Chapter 13 of Fleiss, 1981, for a discussion of interrater agreement). The reliability can be assessed by either having all coders independently code an interview or by having a member of the study staff (or instructor) independently code a fraction of each coder's work. The use of reliability measures will indicate either specific coders who are outliers with respect to the use of the coding

scheme or particular codes which are problematic for one or more coders.

In contrast to the pre-field techniques, behavior coding and the subsequent analysis of the codes requires a sample size sufficient to address analytic requirements. For example, if the questionnaire contains many skip patterns, it will be necessary to select a sample that will permit observation of various movements through the questionnaire. For example, in the discussion of the case study for the Tobacco Use Supplement (Case #3), it was necessary to have purposively sampled smokers, nonsmokers, and prior smokers, to insure observation of all sections of the questionnaire. In addition, the determination of the sample size for behavior coding should take into account relevant population groups for which separate analysis is desirable. Thus, sample sizes for behavior coding may range from fifty cases to as many as several hundred. For this reason, behavior coding is often used in conjunction with field pretests, as a means for providing additional information unavailable from either direct observation or interviewer debriefings. It is also useful to use behavior coding results, if available, during the interviewer debriefings as was done in Leisure Activities Survey (Case Study #2).

Appendix E contains an excerpt from the behavior coding analysis completed for the Tobacco Use Supplement. Although there are no definitive cutpoints for determining whether a

question is problematic, many researchers have used either 10 or 15 percent as an indication that the question is either difficult for the interviewer (question reading behavior) or for the respondent (respondent behaviors). Note that with respect to the respondent's behavior, more than one code can be used for each question. For example, a respondent may first report an inadequate answer, followed by a probe on the part of the interviewer (not coded in the scheme used for this pretest), which finally resulted in an adequate answer on the part of the respondent. Given that the goal is to write unambiguous questions that are both easy for the interviewer to read and that permit the respondent to provide a response without further probing, the above described sequence (if it occurred frequently) would be seen as indicating some problem with the question.

D. Respondent Debriefing

Respondent debriefing involves incorporating follow-up questions in a field test interview to gain a better understanding of how respondents interpret questions asked of them. The technique was originally used many years ago (Belson, 1981), but has gained in popularity in recent years (Fowler and Roman, 1992; Oksenberg, et al, 1991; Esposito, et al, 1991; and Nelson, 1985.) This technique is sometimes referred to in the literature as special probes (Oksenberg, et al, 1991) or frame of reference probing (DeMaio, 1983).

The primary objective of respondent debriefings is to determine whether concepts and questions are understood by respondents in the same way that the survey designers intend. In addition, respondent debriefings can be quite useful in determining the reason for respondent misunderstandings. Sometimes results of respondent debriefing show that a question is superfluous and does not need to be included in the final questionnaire. Conversely, it may be discovered that additional questions need to be included in the final questionnaire in order to better operationalize the concept of interest. Finally, the data may show that concepts or questions cause confusion or misunderstanding as far as the intended meaning is concerned. In any of these cases, changes can be incorporated into the final questionnaire to reduce measurement error.

In contrast to the pre-field techniques, which involve small numbers of cases and qualitative information, respondent debriefing can provide a quantitative method of evaluating problematic areas of the questionnaire. This enables objective documentation of the existence of a problem and its magnitude. In a small-scale pretest, respondent debriefing questions are generally asked of all respondents. For large pretests (and also perhaps for the survey itself), samples of respondents are sometimes administered the debriefing questions. Target populations for debriefing questions are usually determined by specific paths taken during the interview. To reduce respondent

burden and decrease interview costs, a random sample of respondents can be selected. If there are debriefing questions to be asked of a rare segment of the population, it may be necessary to select all households with rare characteristics for the debriefing, to insure that enough cases are obtained to facilitate analysis. It is also desirable to randomly select, within households, the eligible sample person about whom the debriefing questions will be asked. While this may be manageable in an automated interviewing environment, it may be too burdensome for a paper-and-pencil interview. A simpler procedure (but one from which you cannot generalize) is to select the first sample person who meets the target population criteria.

A critical aspect of a successful respondent debriefing is that the right questions must be asked. Question designers and researchers must have a clear idea of potential problems, so good debriefing questions can be developed. Ideas about potential problems can result from the use of pre-field techniques prior to the field test, from analysis of data from a previous round of the survey, from careful review of questionnaires, or from observation of earlier interviews.

When developing the questions themselves, several different options are available. One way is to repeat the response back to the respondent, who is then asked an open-ended question intended to determine how he/she developed his/her response to the question (that is, what kinds of things were included in the

response). For example, in the Leisure Activities Survey (Case Study #2), respondents were initially asked whether they had read any short stories, novels, or poetry in the last year. As part of the respondent debriefing, they were asked to describe the book(s) they read. In this case, interviewers recorded the verbatim responses, which then had to be coded for quantitative analysis. In this example, the content of the reading material was coded to determine whether it fit the definition being asked for (that is, was it fiction or was it nonfiction?). The data indicated that over 20 percent of the books named in the probing questions were works of nonfiction such as histories, biographies or self-help books. These data suggest that there was a significant problem among survey respondents concerning the true definition of a novel (Fowler and Roman, 1992).

Another option is to develop structured closed-ended questions that determine how certain words or phrases are understood and whether the "questions, definitions and instructions proposed for a questionnaire convey the frame of reference desired" (DeMaio, 1983). With closed-ended questions, quantitative data are readily available as soon as the information is keyed. One type of structured question is the vignette, which presents a hypothetical situation and asks respondents how to classify it based on their interpretation of the concept. In the CPS, vignettes were used to determine how respondents were interpreting the concept of work. Respondents

were provided with scenarios describing various work and nonwork activities. The introductions and actual questions asked of the respondent were varied to reflect the different wording of the "work" question in the alternative questionnaires. After the scenarios were read, respondents were asked "Would you report him (her) as working last week, not counting work around the house?", or "Would you report him (her) as working for either pay or profit?". According to Martin and Polivka (1992), the vignette analysis provided information about particular problems of comprehension and interpretation and permitted insights into the conceptual structure that underlies respondents' classifications of different situations. It also "permitted comparisons of the relative restrictiveness or inclusiveness of respondents' interpretations of key survey concepts under different versions of a question" (Martin and Polivka, 1992).

Another kind of structured question asks respondents directly about aspects of the question content that apply to themselves. In the Leisure Activities Survey, additional questions were asked whether they had read specific kinds of reading material (e.g., stories about detectives or spies, fictional stories that appear in magazines) to see whether respondents were including these kinds of material in their original answers.

Respondent debriefing can also be used to determine how respondents interpret and define terms, to determine whether

their definition is consistent with the intent of the survey designers. For example, in alternative questionnaires tested during the redesign of the CPS questionnaire, if a person was a multiple jobholder there were several questions in which persons were asked about the main job and the other job(s) separately. The definition of "main job" was never provided to respondents. To determine if respondents' understanding of "main" job was consistent with the survey definition, a respondent debriefing question was administered in which respondents were asked, "You mentioned earlier that you had more than one job. How did you decide which job was your MAIN job?" Results indicated that only 63 percent of respondents shared the definition of main job as being the "job worked at the most hours" (Esposito, et al, 1991). Based on these results, the definition of main job was included in the redesigned questionnaire so all respondents would know what interpretation was intended.

Due to respondent burden constraints, it is often necessary to restrict respondent debriefing to probe only a few problematic questions or concepts. Otherwise the debriefing becomes too taxing for respondents. As with developing the probing questions, an in-depth knowledge of the survey questionnaire and its problems is necessary to be able to identify which questions should be probed.

The mode of data collection for the survey can also have an impact on the complexity of a respondent debriefing

questionnaire. If a pretest interview is being collected through CATI/CAPI, the specific set of debriefing questions selected for a given respondent can easily be dependent on certain paths taken during the interview. For example, debriefing questions intended to probe respondents' understanding of job search activities can be programmed to show up on the screen only for respondents who report that they are unemployed. Use of a paper-and-pencil questionnaire is much more restrictive with regard to identifying which persons in the household fall into the target group of interest depending on the path of the interview. Less complexity is feasible when previous responses or interviewer check items have to be reviewed to decide whether particular respondent debriefing questions are appropriate.

Respondent debriefings are typically conducted at the end of the interview so the content of the debriefing questions does not bias responses during the interview. (This is especially true if there is to be intensive probing of particular questions.) When making the transition from the pretest questionnaire to the respondent debriefing questions, it is necessary to inform respondents that their role is changing. An introduction to the debriefing questions is desirable. This should tell respondents that additional questions are being asked in an effort to improve the questionnaire, and that they are now acting as informants rather than as respondents.

This change in the focus of the interview is sometimes difficult for interviewers. Respondent debriefing questions attempt to tap into the cognitive processes associated with response formulation, and interviewers are not generally familiar with this approach. As a result, they sometimes find it awkward to ask debriefing questions. To alleviate this discomfort, some time should be spent discussing the respondent debriefing questions and their purpose during interviewer training for the pretest.

Respondent debriefings have the potential to supplement the information from behavior coding (see Oksenberg, et al, 1991). As noted in Section II.C., behavior coding demonstrates the existence of problems but does not always indicate the source of the problems. When designed properly, the results of respondent debriefing can provide information about the sources of the problems. In addition, debriefing data may reveal problems not evident in the response behavior. For example, as mentioned above, the respondent debriefing results indicated that nearly 40 percent of respondents did not share the same definition of "main" job as that intended by the survey designers, and consequently a definitional statement was included in the questionnaire. There was essentially nothing in the behavior coding results that indicated there was a misunderstanding of the concept "main job." Given the large degree of misinterpretation of the concept, it is surprising that the behavior coding data

did not show large percentages of "requests for clarification." However, this is an example of where there is "silent" misinterpretation and unless debriefing probes are asked, the misinterpretation may never be detected by survey designers.

E. Interviewer Debriefing

Frequently, the primary evaluation of questionnaires used in pretests is done through interviewer debriefings. The primary objective of the interviewer debriefing is to provide researchers and study staff with information about weaknesses in the questionnaire wording and/or structure that need to be remedied prior to the questionnaire being used in the field. Unlike other techniques described previously, the main source of evaluation material obtained with this method is the interviewers. They have the most contact with respondents and can enrich the questionnaire designer's understanding of questionnaire problems by sharing the comments provided by respondents during the interview experience. As noted in the introduction, this technique should not be considered sufficient as the sole method of evaluating a pretest, but it is a necessary component of a well-rounded evaluation.

There are a variety of techniques that can be used to obtain information from interviewers about problems with a questionnaire. Interviewers can be debriefed in a group setting, through interviewer rating forms, or through standardized

interviewer questionnaires (also referred to as structured post-interview evaluations). These techniques are frequently used in conjunction with each other. Each of these techniques is described in more detail below.

Group Setting Debriefings

Debriefing interviewers in a group setting similar to a focus group is the most common method used during pretests. Interviewers who conduct the pretest are brought together after the interviewing is completed and asked about their experiences administering the questionnaire. Typically, the moderator of the debriefing will review the questionnaire item by item to ascertain what problems interviewers experienced with regard to question wording, question sequencing and the overall flow of the interview.

It has been suggested that interviewer debriefings consist of no more than 15 interviewers (DeMaio, 1983; Nelson, 1985). This represents a ceiling, taking into consideration all the factors discussed in Section III.B. on focus groups. The staff selected for participation in the debriefing should consist of interviewers with varying years of experience and levels of interviewing skill. This is important because newer interviewers or interviewers who have not acquired good interviewing skills may have different concerns about a questionnaire than experienced or well-skilled interviewers.

A moderator or discussion leader, typically a survey operations staff member or a researcher, guides the discussion using an outline of topics that require attention. The debriefings are tape-recorded so the flow of the discussion is not disrupted whenever a notetaker needs to clarify a point. It has been suggested that a scribe take notes in addition to the audiotaping, in order to have a written summary of the debriefing prior to having the tapes transcribed (DeMaio, 1983; Nelson, 1985).

During interviewer focus groups it is critical that the moderator encourage participation of all attendees. He/she must be tuned in to the group dynamics and insure that a few participants do not dominate. It may be necessary for the moderator to have to continually solicit comments from the more timid participants, to insure that all views are represented.

Group debriefing sessions are generally held within a few days after
interviewing is completed. Other evaluation methods may also be incorporated into the debriefing session to solicit additional information. For example, interviewer rating forms (discussed below) can be administered in conjunction with the group debriefing. Also, if behavior coding has been conducted and the results are available by the time the debriefing session is held, the findings can be discussed with interviewers to learn their opinions about why certain questions were problematic.

It is generally desirable to incorporate some additional methods of obtaining interviewer input, since the primary weaknesses of interviewer debriefings in group settings are that most of the information obtained is of a qualitative nature and "group think" may develop among participants. During testing of new techniques for pretesting, Bischoping (1989) identified several weaknesses associated with the interviewer debriefing technique. She determined that the group debriefing may not identify all problems of a questionnaire, may not produce good estimates of their prevalence, and may lead to question revisions that fail to resolve the most frequently occurring interviewer and respondent difficulties. For example, problems may be identified by one or two interviewers, yet the particular questionnaire item may not be problematic in the overwhelming majority of cases. It may be that a particular difficult or unpleasant incident with a respondent causes an interviewer to overreact with regard to a particular question. As a result of these problems, work is currently underway (Bischoping, 1989; Esposito and Hess, 1992) to develop quantitative methods for use within group debriefing sessions to determine the magnitude of the problems identified by interviewers. For example, once a question is identified as problematic, interviewers are asked to individually indicate what percentage of the time they have difficulty getting an adequate answer to the item.

Another problem associated with the qualitative nature of group sessions is that interviewers' preferences for particular questions may be driven by their desire to keep the interview short and less burdensome for themselves, rather than by their perceptions that the questions are problematic for respondents. During the recent redesign of the CPS, alternative questions about hours worked were tested and interviewers expressed a preference for the shorter series of questions. However, data analysis showed that the longer series of data produced more accurate estimates (Rothgeb and Hess, 1992). Lessler had similar experiences during interviewer debriefings for field testing of alternative questionnaires for a dental supplement to the National Health Interview Survey (Lessler, et al, 1989). Contradictions in results from qualitative data from interviewer debriefings and quantitative information from data analysis demonstrate that other question evaluation methods (e.g., behavior coding, response distribution, respondent debriefings) should be used with pretests to determine how particular questions are working.

While it is recognized that the weaknesses described above exist, there are several strengths associated with the interviewer debriefing technique. As previously stated, interviewers possess an enormous amount of information, given that they are the closest to the interview experience. During the redesign of the CPS questionnaire, information obtained from

interviewer debriefings was instrumental in identifying problematic concepts/terms, problems with the structure of questions, question sequencing, and particular types of questions. Interviewers can also be helpful in identifying solutions to some of the problems and the brainstorming that takes place in the group setting contributes to the valuable suggestions obtained from interviewers. In addition to providing valuable insights with questionnaire problems, the interviewer frequently provide useful information on operational procedures and how they might be improved.

Interviewer Rating Form

Another technique to obtain information from interviewers regarding pretest questionnaires is the use of interviewer rating forms. After the pretest is completed, interviewers complete a standardized rating form and rate each question in the pretest questionnaire on selected characteristics of interest to the researchers. The focus is to quantify the extent of problems, and information about the reasons for the problems is not obtained. The exact content of the rating form can vary according to the needs of the researchers and survey designers. One that has been developed for use at the University of Michigan Survey Research Center (Fowler, 1989) instructed interviewers to rate each question on the following four characteristics:

1. Interviewer has trouble reading the question as written;

2. Respondents don't understand words or ideas in the question;

3. Respondents have difficulties knowing the accurate answer;

4. Respondents have trouble answering in the terms required by the question.

When rating the question, interviewers use the following categories to rate the characteristics described above: no problem evident, possible problem evident, or definite problem evident. In the Leisure Activities Survey (Case Study #2), a modified version of this form was used. It obtained interviewer ratings of three aspects of the question:

1. Interviewer has trouble reading the question as written;

2. Respondents don't understand words or ideas in the question;

3. Respondents have trouble providing answers to the question.

A copy of the rating form used in the case study is included as Appendix F.

Interviewer rating forms can be used in several different ways. First, they can be used as the sole method of collecting information from interviewers. This might be useful when lack of time or resources prohibits the conduct of group debriefing sessions. Alternatively, they can be used in conjunction with group debriefings. When used in this way, they are completed by

interviewers prior to the group debriefing. Interviewers can either be instructed to complete the forms and bring them to the debriefing session, or the forms can be collected beforehand. The decision about which way to proceed depends on the objectives of the evaluators. When interviewers bring the forms to the debriefing session, the forms serve to jog their memories and stimulate discussion about reasons for the problems and/or potential solutions to the problems observed. Having completed the form ahead of time and using it as reference material during the session, the interviewers are more likely to present the broad scope of the problems and not forget anything. However, because the interviewers have access to the rating forms during the debriefing, the two methods of measuring questionnaire problems are not independent. If the questionnaire evaluators are primarily interested in independent measures of the prevalence of problems, then the forms should be collected ahead of time and interviewers should participate in the group debriefing without them. Even with this method of administration, however, it is not clear that the two measures are indeed independent--perhaps filling out the rating forms made some problems more salient to interviewers and subsequently caused the interviewers to discuss them at the group session.

In the Leisure Activities Survey (Case Study #2), interviewer rating forms were collected from the interviewers prior to the group debriefing precisely because independent

measures of questionnaire problems were desired. Results summarized by Fowler and Roman (1992) indicate that all the questions identified as problematic on the rating forms were also identified during the group debriefing session. Fowler and Roman suggest that although the information from the rating forms seems to be redundant with that obtained during the group session, the strength of the rating form may be that it serves as a stimulus to a more thorough debriefing.

Some questions were identified as problematic only at the group session and not on the rating forms. These questions fell into two general categories, question flow problems and definitions of terms (such as whether to include lawn mowing as "gardening done for pleasure"). In the free-flowing discussion in the group, a question was identified as problematic within the context of the larger interview because the flow of the interview was affected by the question, even though the question itself was not problematic.

Standardized Interviewer Debriefing Questionnaire

Standardized interviewer debriefing questionnaires (sometimes referred to as structured post-interview evaluations) are another way of collecting standardized information from interviewers. These debriefing questionnaires differ from the interviewer rating forms described above in that they are a much more flexible tool for collecting information. The interviewer

rating form collects only information about the magnitude of specific kinds of problems related to individual questions. With the standardized interviewer debriefing questionnaire, questions can be included to determine the prevalence of a problem, reasons for the problem, and proposed solutions to a problem (Esposito and Hess, 1992). In addition to providing useful information on questionnaire design issues, standardized debriefing questionnaires can also provide data on the attitudes and behaviors of the interviewers that may affect respondent behavior and consequently responses to survey questions. These data can produce useful information for improving interviewer training or revising interviewing procedures (DeMaio, 1983).

Debriefing questionnaires are designed to be self-administered and they are an extremely cost-efficient way to collect data from all interviewers participating in a field test, as opposed to only 10-15 who can participate in group debriefings. This is particularly advantageous for large-scale pretests, since many smaller field tests can accommodate all interviewers at a single debriefing session.

Both the design of the questionnaire and the structure of the debriefing process are affected by the number of interviewers involved in the field test. If the field test involves a small number of interviewers (less than 15), then the questionnaire can be designed with open-ended questions, since it will be possible for each questionnaire to be reviewed individually. Given the

small number of interviewers and the open-ended design of the questions, the resulting analysis will be qualitative in addition to quantitative. However, if a field test involves a large number of interviewers and each questionnaire will not be individually reviewed, then the debriefing questionnaire should be designed with closed-ended questions, and the data should be entered into a database and quantitatively analyzed.

There are two ways of structuring the debriefing process when standardized debriefing questionnaires are used. First, these questionnaires can be used by themselves, independently of any group debriefing sessions. In this case, the questionnaire is developed by researchers, who determine what survey items are the most problematic and deserving of inclusion in the questionnaire. The debriefing questionnaire can then be administered to all interviewers, regardless of how many there are, or a sample can be selected if desired.

The second way of structuring the debriefing process is appropriate when a large number of interviewers participate in a field test. This is an iterative process, used in conjunction with group debriefing sessions. In this case, group sessions are first used to identify problematic questions or concepts among a small number of interviewers. Then the results are used to develop a standardized debriefing questionnaire that is administered to the entire pretest interviewing staff. Thus, the interviewers participating in the group debriefing make a direct

contribution to the content of the standardized questionnaire. In addition to being useful for field tests involving a large number of interviewers, it is also useful when the field questionnaire is too large to inquire about every survey item. The items identified in the group setting as the most problematic are then included in the debriefing questionnaire for evaluation by all the interviewers.

It is critical that the standardized questionnaire be designed properly since in a self-administered mode, much is left to the interviewer's interpretation. For both open- and closed-ended questions, the questionnaire should include clear instructions to the interviewers so that proper procedures are followed. For example, if interviewers are requested to identify the most problematic item on the questionnaire and their suggestions as to the source of the problem, the questionnaire should clearly state that only one response is allowed.

F. Split Panel Tests

Split panel tests refer to any field test in which more than one version of a question, set of questions, or questionnaire are tested, using replicate samples. The objective of split panel tests for pretesting is to determine which version of the question or questionnaire is "better"; it is therefore critical

to determine a priori a standard by which to judge the different versions. There may also be an analytical purpose--to examine factors which influence responses.

In its simplest form, a split panel test involves the testing of two alternative wordings of a particular question, and all other essential survey conditions remain the same. Such split panel tests have been used to test alternative wordings of questions, alternative response options, and determining whether a "don't know" response should be provided to the respondent. Other versions of split panel tests have involved alternative versions of the questionnaire, for example, changing the order of sections of the questionnaire (to test whether there is a context effect) or alternative administrations of the entire questionnaire (in-person vs. telephone- administered) to examine whether there are mode effects.

Another use of the split panel test involves its function as a benchmark to calibrate the effect of changing questions. This is particularly important in the redesign and testing of a large-scale continuing survey for which the comparability of the data collected over time is an issue. (Comparability of data over time is an issue more generally, but the time and cost involved in benchmarking is usually only justifiable for large-scale surveys.) The final stage in a research program to revise the questionnaire should be to test the final revised questionnaire against the old (control) version before the new questionnaire is

adopted. Comparison of the data from the old and new questionnaires will produce important information about the effects of the questionnaire revision independent of changes over time.

It is important in designing a split panel test to provide for adequate sample size to test differences of substantive interest. Similarly, it is imperative that these tests involve the use of randomized assignment within replicate sample designs, so that differences can be attributed to the question or questionnaire and not to the effects of differential sampling error or incomparable samples. The design of the panels in split panel experiments need to be carefully planned (see Cook and and Campbell, 1979, for a discussion of the design of experiments). In fielding the test, it is crucial to control the randomization and ensure that each panel receives the proper treatment (questionnaire version, mode of administration, etc.) to preserve the integrity of the comparisons. In other respects, administration of a split panel test is the same as for any other field pretest.

Evaluation of split panel tests may include the use of several of the techniques outlined in this section, including comparison of response distributions and examination of item nonresponse (described below), interviewer and respondent debriefing, and comparison of behavior coding results.

Additionally, external or validation data can be used to determine the optimal form of the question or questionnaire.

G. Item Nonresponse and Response Distribution Analysis

The data collected during a field test are an important source of information about how well the questionnaire works. Two types of data analysis are useful: item nonresponse rates and response distributions. Analysis of pretest data is generally done using raw data (unedited and unimputed), because the focus is on how well the respondent understands and responds to the questions. Editing and imputation improve the quality of the data and thereby have the potential to mask questionnaire problems.

Item nonresponse rates are defined as the percentage of persons eligible for a question who do not provide a substantive response. Examination of item nonresponse rates is useful for all field pretests, and the results can be informative in two ways:

1) "don't know" rates can be examined to determine the extent to which a task is too difficult for respondents to do; and,

2) refusal rates can be examined to determine the extent to which respondents find certain versions of a question to be more sensitive than another version.

Response distributions display the frequencies with which answers are given by respondents during the pretest. They are

calculated on the basis of eligible persons who respond to a specific question, excluding those for whom a "don't know" or "refusal" is obtained. Evaluation of the response distributions for survey items can often help to determine if different question wording or question sequencing produces different response patterns. (This is especially useful if more than one version of a question or questionnaire is being tested, such as in split panel tests.) Specifically, Rothgeb and Hess (1992) found during the redesign of the CPS that response distribution analysis can be useful in evaluating questionnaires in the following ways:

- to determine the impact of direct questions versus volunteered responses (e.g., asking directly about unpaid work in a family business instead of assuming respondents will know to voluntarily report such work);

- to determine the impact of using different methodologies to obtain the same information (e.g., using different strategies to acquire information on the number of hours a person worked last week);

- to determine the impact of question sequencing (e.g., using different question ordering to determine if response patterns are affected); and,

- to determine if alternative questions are more inclusive of the targeted universe (e.g., rewording a question so persons who may be only marginally in the targeted universe get captured).

A constraint of response distribution analysis is that it is most useful when pretesting either more than one version of a questionnaire or a single questionnaire for which some known

distribution of characteristics exists for comparative purposes. As such, it requires relatively large sample sizes to achieve statistical significance between versions. If only one version of a questionnaire is being tested and there is no reliable known distribution of characteristics, the benefits obtained from response distribution analysis are greatly reduced.

When using response distribution analysis for question evaluation in split panel tests, the results do not necessarily reveal whether one version of a question produces a better understanding of what is being asked than another. Knowledge of differences in response patterns alone is not sufficient to decide which question best conveys the concept of interest. In addition, sometimes response distribution analysis demonstrates that revised question wording has no effect on estimates. For these reasons, response distribution analysis should not be a sole method for evaluating modifications in question wording or item sequencing. It is useful only in conjunction with other question evaluation methods such as respondent debriefings, interviewer debriefings and behavior coding.

IV. Considerations in Developing Pretest Plan

There is no one "right" way to conduct a pretest. Depending on the objectives of the testing activity and the amount of time and money available, there are a number of options. In this

section of the report, we present some of the practical considerations that must be addressed in developing a pretest plan that makes the best use of the techniques discussed above.

A. Time and Cost

It is very difficult to give actual time and cost estimates for the methods described in Section III, because of the variety of ways in which they can be implemented. In this section, we provide a discussion of the factors that need to be considered in estimating time and costs for any particular testing activity. We also provide ballpark estimates of the amount of time required. The estimates are presented in terms of person-time; that is, time devoted by staff to complete the tasks. The total elapsed time may be longer if staff members are splitting their time among different tasks. These estimates are very crude because the amount of time can vary greatly depending on several factors: the amount of developmental time required to get materials ready for use in the field or laboratory, the length and complexity of the questionnaire being tested, the experience of the staff assigned to the project, the amount of time available for the project, and the amount of coordination required between the sponsor and divisions within the Census Bureau.

Ballpark estimates of cost are not included because much of the cost derives from salary costs, these vary greatly depending on who does the work, and they also change over time.

The pre-field testing methods described in Sections III A. and B. are relatively cheap compared to the cost of a field test. The time involved varies depending on the method used.

Cognitive interviews. Cognitive interviews are generally more time-consuming than focus groups, but the time invested has payoffs in terms of the depth of the information collected from each respondent. The time required involves time for planning the research (how many respondents will be interviewed and what kinds of characteristics they should possess [e.g., smokers vs. nonsmokers, never-married vs. divorced mothers]), recruiting respondents, developing a protocol for probing of respondents about suspected or known problems in the questionnaire, conducting the interviews, preparing summaries of each individual interview, and preparing a report that summarizes problems and makes recommendations for changes to the questionnaire. The amount of time required varies according to how many interviews are to be conducted, and how many interviewers and how much of their time is available. The general range is two to four months. This process can be shortened if individual interview summaries are not prepared; however, we do not recommend that approach because the summaries are useful both for helping the

interviewers to get more insights about the response process and for convincing the sponsors of the problems with the questions.

The cost of conducting cognitive interviews is mostly comprised of the salaries of the staff conducting the interviews. Two other cost factors are also relevant. The first is the cost of small reimbursements to interview participants for their time and travel expenses. The second is travel costs to the site of the interviews if they are not conducted in the CSMR Response Research Laboratory. Depending on the objectives of the research, regional differences in the quality of response to a questionnaire may mandate that interviews be conducted at sites across the country. Additionally, local travel reimbursements may be appropriate if circumstances require that interviewers travel to participants' homes or offices to conduct the interviews. This may be the case if, for instance, the survey requires school principals to look up information such as the number of students in the school or if the questionnaire requests people to report the brand names of products they have eaten.

Focus groups. Focus groups are the quickest way of collecting information from respondents, but as noted in Section III, they are not as useful as cognitive interviews for questionnaire pretesting purposes. Because respondents are assembled as a group rather than contacted individually, the time spent has a larger payoff in terms of the amount of information obtained. The time required involves time for planning the

research (how many groups will be conducted and what kinds of respondents will be needed), going through the procurement process (if the groups are contracted out), recruiting respondents, developing a topic outline for use for use in guiding the discussion, conducting the focus groups, transcribing tapes, and writing a report. Depending on the complexity of the project, this process can take anywhere from three weeks to three months.

Focus groups can either be conducted in-house or by an outside contractor, and the cost involved varies depending on which approach is used. Availability of funds and availability of in-house personnel to conduct the focus groups are the main factors that need to be taken into consideration when making a decision about how to conduct the groups. Contracting for focus groups generally costs approximately $4000 per group, and includes respondent recruitment, rental of space, preparing a moderator's guide, conducting the focus groups, and preparation of a summary report. Generally, preparation of an additional report focusing on the implications of the results for questionnaire testing purposes is appropriately undertaken in-house.

If the groups are conducted in-house, the cost factors involved are staff salaries, respondent reimbursements, and, depending on where the groups are conducted, travel and renting of space and equipment.

Behavior coding, respondent debriefing, and interviewer debriefing are all methods that are incorporated into a field test. For these methods, the biggest cost and time factor by far is the field test itself. Compared to that expense, the other factors involved are minimal. The following presentation discusses the time and cost factors over and above those involved in a field test.

Behavior Coding. The amount of time required to incorporate behavior coding into a field test is very small, since the data are collected while the field test is proceeding. A coding scheme needs to be developed and coders need to be trained. (This can be done simultaneously with other activities to prepare for the field test.) After the data collection ends, the tapes must be coded, and the data entered into a computer database and analyzed so a report documenting the results can be prepared. The amount of time required for these activities ranges from two weeks to two months, depending on several factors. If the coding is done live, either in the field or at a CATI facility, then it will be completed at the same time as the interviewing is completed. If the interviews are taped and the coding is done after the fact, it still need not take too much longer, although other factors need to be considered. If only a sample of interviews are to be coded, additional time may be required to sample the tapes. Time required for coding (and subsequent analysis) will also depend on the length of the questionnaire

being tested, the complexity of the coding scheme, the number of available coders, the number of cases being coded, and the logistics of getting the tapes to the coders. (Typically, one can expect the time to code an interview to be three times the length of time to conduct the interview.) If all the interviews are to be coded and logistics of getting the tapes to the coders can be arranged, the coding can be done on a flow basis and it can be completed within a few days of completion of the interviewing. Programming a data entry system can be completed before the coding begins, and data entry time can vary greatly depending on the number of questionnaires and the number of coders' verbatim comments that need to be entered. Data entry time can be shortened if the coders listen to the tapes and key the data directly into the computer. Analysis and preparation of tables summarizing results of individual questions and writing up a report that documents the results can all be completed within four weeks.

When planning the field test itself, one other thing needs to be kept in mind prior to the data collection. If interviews will be conducted by telephone, permission must be granted from the Department of Commerce to tape the interviews for later coding. This involves writing a memorandum for approval by the Chief Financial Officer and Assistant Secretary for Administration. (A copy of the memorandum used for the Tobacco Use Supplement [Case Study #3] is included as Appendix G.) Other

preparations can be made simultaneously, but approximately five weeks needs to be allowed to secure the necessary approval.

In terms of costs, the major factor is salary costs for the coders, the data entry clerks and the analysts. These costs can vary, depending on the type of personnel used. As noted in Section III, the coders can be professional staff who are involved in the questionnaire development, or clerks or field supervisors whose only involvement in the project is the behavior coding. Data entry is generally done by clerks, and analysis is done by the staff involved in the questionnaire development.

There are some start-up costs for behavior coding that involve purchase of tape recording equipment (if the coding is done from tapes), and telephone connectors (if interviewing is done by telephone). An ongoing cost for any behavior coding operation is the cost of tapes. All in all, these costs are very small compared to the cost of conducting the field test.

Respondent Debriefing. Respondent debriefings are conducted immediately after the interviews themselves, while the interviewer and respondent are still in contact or immediately after the respondent completes a self-administered form (either paper-and-pencil or automated). As a result, the evaluation data are collected simultaneously with the questionnaire data. Preparation activities for the respondent debriefing can be carried out while other preparations are made. These include the critical task of developing the respondent debriefing questions,

which can be time-consuming, and getting the questionnaires ready for the field. If the respondent debriefing occurs as part of CATI/CAPI/CSAQ data collection, the questionnaire must be programmed. In addition, specifications for sampling respondents or selecting which respondents get asked which questions may also need to be developed and programmed. After data collection is complete, additional time is required to enter the data, analyze it, and write a report. (Data entry may be completed as part of the main survey, or may be done during data collection if CATI/CAPI/CSAQ is used.) Because respondent debriefing questionnaires are typically short, the time required for these activities is also short. The additional time required to incorporate respondent debriefing into a field test ranges from two to four weeks. The additional costs involved, beyond those for the field test, are small expenses for data entry and computer time for analysis, with the bulk being allocated to staff time for analysis and report-writing.

Interviewer debriefing. Interviewer debriefings are typically conducted shortly after data collection for a field test is completed. After the debriefing session is held, the report can generally be prepared within one or two weeks. If interviewer rating forms or standardized interviewer debriefing questionnaires are used in conjunction with interviewer debriefing, some additional time may be required both prior to the debriefing to develop the forms used, and after the

debriefing to analyze and integrate the quantitative data. The number of cases involved is usually small, and even in this case, the debriefing report can generally be completed within two to three weeks.

The costs involved in conducting interviewer debriefing sessions are relatively small. They include the cost of tapes for tape recording the session, additional salary and travel costs for the interviewers to attend the session, and salary costs for staff to conduct the session, review the tapes and compile the debriefing report. If the field test is conducted outside of the local DC area, additional travel costs may be incurred for staff at attend the debriefing (however, in this case the travel cost specifically associated with the debriefing would be small compared to the cost of travel associated with the field test as a whole). For a large-scale field test that uses standardized interviewer debriefing questionnaires, additional costs for data keying and computer analysis may be incurred.

Split Panel Tests. As noted previously, split panel tests are a specific kind of field test in which alternative versions of a questionnaire or survey procedures are used and the results compared. Again, the major time component is the amount of time required for planning and conducting the data collection, the same as for any field test. Small amounts of additional time (perhaps a week) may be required to develop the alternative version of the questionnaire or procedures. This does not

include the time for forms design and printing already required as part of the field test. The most time-consuming addition involves analysis of the results. By its nature, an experimental design stresses the content of the data collected, and places ultimate importance on analyzing the performance of the various alternatives. Time required for this task will be discussed in more detail below.

There are several additional cost components for a split panel test. Depending on the nature and complexity of the alternatives, these may include additional costs for forms design and printing, for development of a CATI/CAPI instrument, for development of training materials, and also for expanding the sample size. Additionally, there will be costs associated with programming and conducting the data analysis and documenting the results.

Analysis of Item Nonresponse Rates and Response Distributions. In contrast to the other methods discussed thus far which involve collection of information separate from or complementary to the field test, this is a technique for evaluating the data that are collected either during a field test or a split panel test. The time factors that need to be considered include time for data capture, file preparation, file documentation, analysis, and preparing a report of results. Time for data capture may differ depending on whether the questionnaire is developed for processing through CATI, CAPI,

CSAQ, FOSDIC, or keying, and depending on the number of questionnaires included in the field test. Generally, data capture should be completed within four weeks after the end of data collection. File preparation also involves different amounts of effort according to how the data were captured, and should be complete within five to seven days after data capture is complete. Time required for analysis and report-writing depends on the objectives of the field test, and on whether custom programs are developed or packaged software programs are used. Simple tallies of item nonresponse and response distributions for each item should be completed, along with a brief report, within one to two weeks of delivery of a final data file, depending on the length of the questionnaire. The analysis of split panel tests with alternative versions of questionnaires may be more complicated and take longer. In this case, results should be complete within three to six weeks of delivery of a final data file.

The cost factors to be considered include computer costs and staff time for programming, analysis, and report-writing. Obviously, the more complicated and time-consuming the analysis, the more costs will be incurred at this stage of the project.

B. OMB Clearance

One of the realities that must be confronted in developing a pretest plan is the requirement for clearance by the Office of

Management and Budget (OMB) if more than nine individuals are asked to complete survey forms.

The process of obtaining OMB clearance is time-consuming. A clearance package (consisting of an SF-83, "Supporting Statement," and copies of all forms and materials) must be prepared (usually by the Census Bureau with assistance from the sponsoring agency) and approved through both the Census Bureau and the sponsoring agency before it is submitted for approval at the departmental level. In the Department of Commerce, 30 days must be allowed for departmental approval; at other agencies, the time required for this approval can range up to three months. Only after this approval is secured can the final package go to OMB. Unless it is an emergency submission, OMB must be allowed an additional 90 days to review the package and provide approval/denial. Thus, a minimum of four months generally needs to be built into the time schedule just to accommodate OMB clearance requirements.

In some situations, this process may be simplified. CSMR has obtained from OMB a generic clearance that applies specifically to questionnaire pretesting research. In specific situations, the extended waiting period can be eliminated from time schedules for pretesting questionnaires. The circumstances under which the generic clearance applies are as follows:

1) the clearance involves pretesting of survey questionnaires or procedures, and the testing

will yield information that will be useful for making changes;

2) the methodology for the pretest activity consists of any of the following:

- cognitive interviews;
- focus groups;
- respondent debriefing questionnaires;
- field tests in which interviewer/respondent interaction is coded; or
- split sample experiments;

3) the number of cases included in the pretest activity is small (no more than a few hundred);

4) the testing does not involve a "dress rehearsal" of survey procedures;

5) there is no experimentation with incentives; and

6) approval to use the generic clearance is granted by CSMR, which is responsible for overseeing the clearance and monitoring the respondent burden hours.

If these criteria are met, the current procedures stipulate that OMB be notified by letter at least one week in advance of the pretesting activity. The letter is written by Census Bureau staff responsible for conducting the activity, for review and signature by staff in CSMR. The letter contains the following information:

- the name of the survey for which the testing is being conducted, and the sponsor;
- description of the test procedures;
- kind of testing (cognitive interviews, focus group, field test, etc.);

- location and dates of data collection;
- number of interviews to be conducted;
- plans for selecting respondents;
- data collection procedures;
- assurance that data are confidential and voluntary;
- methods to maximize response rates;
- respondent burden hours;
- names and addresses of contacts for statistical aspects and data collection.

This is the same basic information that is required in an OMB supporting statement, although in less detail. As with a regular OMB submission, informational copies of questionnaires, debriefing materials, and other materials sent to respondents must accompany the letter.

Another requirement of the clearance is that the results of the pretesting activity must be written up and sent to CSMR. This is because OMB requires that the Census Bureau annually submit a report that documents the work done under the clearance and the results of the testing.

The existence of this generic clearance greatly facilitates the Census Bureau's ability to pretest questionnaires, since the amount of time required in advance of testing is drastically reduced. The time constraints basically consist of the ones outlined in the previous section, without any additional lag time. If time is a luxury in the planning of a particular

pretest and more than one round of testing is conducted (using more than one of the above techniques), a separate letter must be sent to OMB describing each activity.

The generic clearance is not a panacea for all questionnaire testing, since there are situations in which large field tests are necessary to meet the objectives of the pretest. In those cases, the routine and time-consuming procedures for obtaining OMB clearance must be complied with. However, for other smaller-scale testing needs, the generic clearance provides a less cumbersome way to increase the amount of testing of Census Bureau questionnaires and improve the quality of data collected.

C. Study Design

It is impossible in a protocol manual of this nature to prescribe specifications for the conduct of a particular pretest. Rather, the intent of this section is to enumerate the issues, somewhat like a checklist, that should be considered in the design of a pretest or a set of pretesting activities. Among the considerations should be:

- the nature of the pretest, that is, the objectives of the pretest and the set of techniques (as discussed in Section III) that will be used to meet those objectives;
- the design of the pretest sample, including the sample size, the selection of the sample, and respondent recruitment;
- the mode of pretest data collection (self-administered, face-to-face, or telephone);

- the medium used to capture the pretest data (paper-and-pencil or computer-assisted data collection);
- the content of the pretest questionnaire;
- the setting for the pretest data collection (respondent's home, laboratory facility, etc.);
- the length of the study (that is, the number of weeks available for conducting the interviews);
- the number of interviewers required; and
- respondent selection rules, including self and proxy decisions, number of respondents per household, etc.

Each of these issues will be discussed briefly below.

Sample design. Several issues related to the sample design are critical to the design of the pretest. A clear statement of the test objectives will help in making decisions concerning these issues. They include:

- Type of sample: Do you want to conduct a pretest with a nationally representative sample or is a purposive sample or a sample selected from a limited number of sites sufficient? As with most issues that will be enumerated in this section, the answer to this question is, in part, determined by the techniques used to conduct the pretest. In using cognitive interviewing techniques, most researchers begin with a small number of informants, selected from a limited number of sites, focusing on the population group or groups of primary interest. In contrast, split panel tests, by their very nature, require the use of random assignment to treatments. This can be achieved by random

assignment to statistically representative samples or to convenience samples.

- Sample size: The first question of interest in determining sample size is whether the data collection effort will result in statistical analyses or whether the data are seen as providing qualitative information. Once again, the use of many of the techniques, including cognitive interviews, respondent and interviewer debriefing, and focus groups, are not dependent upon large sample sizes for conducting statistical tests. Behavior coding can prove to be informative, even when applied to a relatively small sample. However, if the tests involve the comparison of different forms of the question or questionnaire or an analysis of item nonresponse rates, or different behaviors associated with alternative questionnaire designs, it is necessary to determine the analytic goals, determine what size effect is desirable to detect, and from there, set the sample sizes to provide sufficient sample size and power to meet those goals.

- Sample selection and recruiting: There are a variety of ways to select a sample. These include the use of expired sample for a major study, so as to facilitate the selection of individuals with particular characteristics (e.g., selecting from the outgoing rotation of the CPS), random digit dialing (for telephone-administered questionnaires), area probability sampling, recruiting informants through

advertisements and posting, mall intercept samples, etc. Once again, the determination of the sample selection is to a large extent based on the goals and techniques used in a particular pretest.

Mode of Data Collection. The issue of mode of data collection is not a question for many of the techniques enumerated in Section III. For example, the use of cognitive interviewing techniques and focus groups demands face-to-face interaction, although both techniques have been used for observing and discussing self-administered questionnaires. Ideally, a field test should reflect the mode or modes that will be used in the actual study, to eliminate the potential confounding of the effects of mode on the findings. Although the literature on mode effects indicates mixed findings, we do know that the nature of the interaction is different across the various modes. In the interest of saving costs, however, pretesting of personal visit questionnaires may be conducted by telephone from a centralized facility. When changes are made in the mode of data collection, it must be recognized that differences in the dynamics of the interaction between the respondent and interviewer may affect the test results. Self-administered vs. interviewer-administered questionnaires raise different issues for the questionnaire designer. For example, the need to have a document that is visually pleasing with simple

skip instructions may be a more important criterion for the former.

Medium for Data Collection. For the purposes of this manual, "medium" refers to the means by which the data are captured--either paper-and-pencil or computer-assisted interviewing. Similar to the mode of data collection, it is useful to use the medium of interest for the main study in designing a pretest, so the dynamics of the actual response process will remain constant. Also similar to mode of data collection, some of the techniques outlined in Section III are not conducive to using computer-assisted interviewing (e.g., cognitive interviewing techniques). However, one can imagine using either of these media in conjunction with many of the other techniques discussed in Section III.

In determining whether to use paper-and-pencil or computer-assisted methodology, one critical element to consider is the length of time available in the planning phase of the pretest. Traditionally, the use of computer- assisted interviewing techniques has required longer lead times for "authoring" an instrument (around six months), especially when compared to pretest questionnaires which are often xeroxed rather than printed. Content of the Pretest. As noted in Section II, the objective of a pretest can be quite variable, ranging from the revision of a single question to the design of an entire instrument. All of the case studies discussed in Section V

represent tests of supplements to an ongoing survey--either to the Current Population Survey or the National Crime Survey. In conducting a pretest of a supplement, one needs to decide whether it is important (from the viewpoint of content or respondent burden) to administer the main survey prior to testing the revised supplement. Similarly, for any redesign effort short of a complete questionnaire, one needs to determine whether testing can be limited to the revised section(s) of the instrument or whether the entire questionnaire needs to be administered. In making this decision, the possibility of context effects (Schuman and Presser, 1984) needs to be considered. The content of previous questions may affect how respondents answer subsequent questions. Therefore, deleting questions on a pretest questionnaire may create differences in how respondents understand a later question in the pretest and in the actual survey. Using some of the more exploratory techniques, for example, cognitive interviewing or focus groups, it may be beneficial to limit the research to the items of particular interest. However, with respect to larger field tests using behavior coding, debriefing of interviewers or respondents, or a split sample experiment, it is generally useful to test the entire questionnaire. The tradeoff, of course, is whether the amount of resources (both time and money) are available for testing such an extensive design.

There is a literature that documents the kinds of effects that can occur with respect to changes in the wording and sequence of questions, length of reference periods, presentation of response categories, etc. It may be useful to refer to previous research (e.g., Belson, 1981; Bradburn and Sudman, 1979; Biemer, et al, 1991; Converse and Presser, 1986; Payne, 1951, Schuman and Presser, 1981; Fowler, 1988; Sudman and Bradburn, 1974; Sudman and Bradburn, 1982; Turner and Martin, 1984) in developing the content of the pretest questionnaire.

Another issue related to the content of the questionnaire being tested involves whether the questions are part of a time series. A balance between the measurement error affecting the questions and the desire to maintain comparability over time needs to be struck. It may mean that question wording should be kept intact even though it is not ideal, or that some parts of a questionnaire should be kept intact to maintain key trends.

Setting. Where should the pretesting activities be conducted? Once again, this decision is driven by the types of techniques used. Cognitive interviewing can be conducted anywhere it is possible to tape record or otherwise record (in detail) respondents' answers, but many researchers conduct such interviews at a laboratory facility containing audio and visual taping capabilities and a one-way mirror to permit unobtrusive observation of the interviews. However, the respondent's home or office may be a better choice if the questionnaire being tested

requires the use of records to answer the questions. Focus group research needs to be conducted in a neutral facility, which can accommodate the moderator and the informants easily and which permits taping of the session. Field tests involving other techniques such as behavior coding, respondent debriefing, or an eventual analysis of a split sample experiment or response distributions can be conducted in the respondents' homes, from a centralized telephone facility, or by telephone from interviewers' homes. The advantage of a centralized telephone facility is the ease with which interviews can be monitored (that is having a third party observe), thereby facilitating the recording of behavior codes or other observational data. An advantage of automation is that it provides much more control in randomizing the treatments in a split panel test and also allows more complex designs.

Field Period and Number of Interviewers. The size of the sample, the design of the sample (that is, the number of sites in which the pretest is being conducted) and the length of the questionnaire all affect the tradeoff between the length of the field period and the number of interviewers. Ideally, one needs to have enough interviewers so the findings of the pretest are not confounded with interviewer effects. With only one or two interviewers, the results may reflect the effect of the particular interviewer rather than the effect of the instrument being tested. However, each additional interviewer adds

additional fixed costs, for example in training and supervising. One also needs to consider the length of time available for conducting the pretest and the most efficient use of that time. For example, if two months are available for fielding the pretest, it may be more efficient to use six interviewers and complete the work in one month rather than use three interviewers who would need to work the entire two months. As with most of the design issues raised in this section, decisions about length of the data collection period and number of interviewers are highly dependent on several of the other design features, including sample size, mode and medium for data collection, etc. In determining field period and number of interviewers, one must also decide on the level of effort that will be expended to obtain response from a particular sample unit. For example, how many callbacks will be required to interview reluctant respondents or respondents who are not at home? Will refusal conversion be part of the pretest operations?

Respondent Rules. Finally, who should the respondent for the pretesting activities be? The answer depends to some extent on the type of pretesting activity. Focus groups and respondent debriefing require self-response, since the focus is on how the respondent interprets and answers the survey questions. Cognitive interviews can be structured to use self-response or proxy-response, but they should not involve both with the same respondent. Exposure to either the concurrent or retrospective

think aloud method would contaminate the effectiveness of the technique the second time through the questionnaire, because the respondent will not be hearing, understanding, and thinking about the question for the first time (in a concurrent think aloud) and in a retrospective think aloud, recall of self- and proxy responses can be confused. In practice, cognitive interviews are generally conducted using self-response. For field tests, the respondent rules for the actual study are an important determinant. If the actual study permits a mixture of self- and proxy-responses, then the pretest will provide more useful information if both types of respondents are included.

D. Other Issues

In developing a program of pretesting activities, there are several other general issues that need to be addressed. Some are applicable regardless of the methodology used, while others are relevant only for field tests. These issues are as follows:

- coordination among divisions and sponsors;
- preparation of pretest questionnaire and related materials;
- training; and
- data processing.

This section briefly describes each issue and discusses its application in the pretesting process.

Coordination among divisions and sponsors. In planning and executing a set of pretest activities, there are many different actors who perform different functions. In order for the pretest to be successful, good communication and proper coordination among all the responsible parties are essential. It is also extremely important that a project plan be developed which outlines the responsibilities of each involved division and associated deadlines/due dates for the various assigned duties.

For the Census Bureau demographic survey program, Demographic Surveys Division (that is, the staff responsible for representing the sponsor on survey content and timing) serves the role of general coordinator to ensure that pretest goals and objectives are met. It obtains commitments for staff resources from other divisions and coordinates the schedules and activities of all responsible parties. It also serves as the Census liaison between the sponsor and other divisions within Census.

Depending on the type of pretest activity, other divisions may also get involved. If pre-field activities such as focus groups or cognitive interviews are conducted, participation of CSMR staff may be involved, either for training subject matter staff or for conducting the activities. If a field test is conducted, several other divisions may also participate, depending on the size and formality of the test. Demographic Statistical Methods Division may get involved with selecting the sample--its size, design, source, etc. Administrative and

Publication Services Division (APSD) may be recruited to turn rough draft questionnaires and other documents into camera-ready forms and manuals, and to facilitate timely and high-quality printing of materials. Field Division coordinates the work of the interviewers (who are officially known as Field Representatives) and all regional office and telephone center staffs in conducting the pretest, as well as coordinating and/or writing training and office instructions. Data Preparation Division in Jeffersonville, Indiana, may be requested to perform clerical or keying duties, maintain supplies of forms, or ship printed forms to regional offices. Good coordination among all parties involved, including the sponsor is crucial. When schedule changes or problems occur, all divisions who are involved in the activity should be informed. When decisions about the final content of the questionnaire are made, the comments of all staff who have been involved in the questionnaire development should be solicited.

Preparation of Pretest Questionnaire and Related Materials. For any kind of pretesting activity, the focus is on the questionnaire under development. Thus, there must be an instrument for use during the testing. Depending on the type of activity, the questionnaire may be in various stages of readiness. In cognitive interviews, for example, a typed and xeroxed version of the questionnaire is sufficient. Cut-and-paste revisions of previously-used questionnaires can be given to

respondents in self-administered cognitive interviews. For a field test, however, camera-ready versions of the questionnaire are generally designed by APSD, proofed by the subject matter division, and printed. If timing is an issue, the sponsor may elect to contract out the design and printing of forms and other materials. Interviewer manuals and other field materials are generally developed by either DSD or Field Division, and either xeroxed or printed, as timing or quantities require.

Training. An essential ingredient to the success of a field test is good training. The people who actually conduct the interviews must have a fundamental understanding of every facet of the survey. Interviewer training packages for a pretest are prepared either by subject matter experts in the participating divisions or by the Training Branch of Field Division, working closely with subject matter experts. The route that is chosen for development of training materials generally depends on time constraints and availability of staff. Training may be administered through self-studies that the interviewers are asked to complete at home. The self-studies introduce the interviewers to the basic survey concepts and procedures. Training may also be conducted in a classroom setting, whereby groups of interviewers are trained by the survey supervisor in Field Division. Classroom training is usually provided through a "verbatim training guide" read by the trainer and practice interviews. This mimics actual survey procedures, in which

interviewers in all geographic areas must receive exactly the same training. Another training option that is sometimes, but not too frequently, used is less structured classroom training, which is guided by just an outline or brief description of what is to be covered.

These two techniques are often used in combination--a self-study to introduce the survey concepts and classroom training to provide in-depth preparation on survey specifics such as skip patterns and question wording. An important function of classroom training is to give interviewers a chance to complete practice interviews, to get a feel for the flow of the questionnaire and get familiar with its content.

Data Processing. The completed survey questionnaires need to be processed before any analysis can begin. At a minimum, processing includes
clerically checking in and reviewing the forms, and keying the data. At the pretest stage, the sponsor may elect to do very little additional clerical or computer data cleaning, and may decide not to weight the data, especially if they are not going to be published. In that case, the sponsor may want to see the data presented exactly as collected, in order to better identify questions that do not yield the expected responses. Inconsistencies are often "flagged" at the pretest stage, but the data are left intact.

In some instances (for example, in a split panel test) more complicated processing may be warranted. Then, there are two additional layers of processing: computer "cleaning" the data through consistency checks, editing and/or imputation, and weighting the sample data (when required) to reflect a national universe. These activities create a data file that can be used to compare estimates across panels or to national totals for detailed substantive analysis.

E. Reporting of Results

Reporting the results of any pretesting activity is crucial to getting the most out of the project. In the descriptions of many of the techniques described in Section III, summaries and reports are included as part of the standard procedures. These reports serve several purposes: 1) they force the person preparing them to concentrate on the interview, focus group, etc., and think about the implications of the interaction for the revision of the questionnaire; 2) they constitute a common body of knowledge that all the staff working on the project can tap to revise the questionnaire; and 3) they provide documentation of the problems with the questionnaire for survey sponsors.

One of the techniques described in Section III, analysis of item nonresponse rates and response distributions, is a method for reporting the results of a field test. This stage of the

field test is often neglected entirely within the Census Bureau; the data are turned over to the sponsor for substantive examination. However, analysis of these relatively simple aspects of the data is important even for subject matter specialists here. This facet of knowledge about how the questionnaire worked is important for joint discussions about making revisions to the questionnaire.

There is always a temptation to eliminate the reports because they are too time-consuming; however, this is time well spent in terms of its impact on the quality of the final product, the revised questionnaire. Furthermore, by documenting the results, they will be available for other staff working on related questionnaires or later rounds of a survey, so that people will not have to keep re-inventing the wheel.

After the results of the questionnaire pretesting research are reported, recommendations for changes to the questionnaire have been suggested, and revisions have been made, it is useful to test the questionnaire again. This testing checks to see if the changes that have been made address the original problems without creating new ones. This can either be done through a round of cognitive interviews or by means of another small field test.

F. Implementation Plans for the Survey

Having completed all the work of planning, conducting, and reporting the results of a program of pretesting activities, it is important to ensure that the outcomes of the research are integrated into the plans for the actual survey. Pretest results sometimes sit around and collect dust, both at the Census Bureau and at the sponsoring agency. In order to collect survey data of the highest quality possible, it is imperative that the results of the research actually get incorporated into the arrangements for the survey. When decisions about the questionnaire and survey procedures are made, input should be sought from the staff who have been involved at the questionnaire development and testing stages. While there may be operational reasons why some recommendations cannot be accommodated, discussion among all the parties involved may come up with additional suggestions that both address problems and are feasible within operational constraints.

V. Case Studies

As part of the process of developing this pretesting protocol, three demonstrations of the use of expanded pretesting activities were conducted. These case studies were chosen to represent a range of situations as far as the length of the questionnaire, time, and available funds are concerned. In this section, we present descriptions of the pretesting activities

conducted as part of each of these demonstration projects, along with some of the results.

A. Short Time Frame, Few Resources: The CPS Supplement on Child Support and Alimony

Background. The Office of Child Support Enforcement, U.S. Department of Health and Human Services sponsors a supplement to the Current Population Survey on a biannual basis, during April. The data from the study provide estimates of the population of children "at risk" of receiving child support, that is, children for whom one or both parents are not residing with the child. During the 1988 and 1990 administrations of the survey, two problems became evident. First, the screening questions used to identify individuals providing care to a child at risk of receiving support improperly excluded all custodial fathers and a subset of custodial mothers. In addition, the survey was plagued with high levels of item nonresponse. Although there were several other concerns with question wording and the administration of the supplement, the two items enumerated above were of highest priority for redesign efforts.

Constraints. Two major obstacles precluded the design of a large research or field effort. First, timing was quite short. The sponsors of the survey made known their concerns in December, 1990; final design of the questionnaire needed to be completed by August, 1991. Thus, it would have been difficult to design a study and permit adequate time for OMB clearance, conduct the

study, and complete analysis prior to the August deadline. (The Census Bureau did not receive authorization for its generic pretest clearance until September, 1991). Second, only a small budget had been allocated for pretesting and developmental activities.

Pretest Design. In light of these limitations, a study plan was developed to: (1) examine possible reasons for the high item nonresponse rates; (2) test alternative question wording that would permit the identification of the entire population at risk of receiving child support; and (3) provide guidance for developing an ongoing research plan to address other issues within a longer time frame.

The research was conducted in two stages. During the spring of 1991, the Census Bureau conducted cognitive interviews with respondents using the 1990 version of the child support enforcement questionnaire as well as a revised version of the questionnaire introduction, which identified eligible respondents. Respondents were recruited from Washington, D.C. and the suburban Maryland area through advertisements that were posted in local child support offices, grocery stores, libraries, and other public places. Contacts were also made with officers of local chapters of Parents without Partners and advertisements were placed in local newspapers. Seven informants who varied with respect to race, education, age, gender, and employment status were recruited.

The second stage of pretesting focused on a particular subset of the population--mothers who have children eligible for support from more than one father. This group of custodial parents, for whom the questionnaire was known to be especially difficult, was of special interest to the sponsors of the survey. Eight additional interviews were conducted with informants who were recruited from a local child support office.

Techniques. As noted above, the primary goals of the cognitive interviews were twofold: to test a new questionnaire introduction designed to capture the universe of all children/parents at risk of receiving child support and to understand the reasons for high item nonresponse rates. In addition, cognitive interviewing techniques were used to understand the respondents' interpretations of several questions. The majority of the interviews were conducted using a concurrent think aloud format, where respondents were asked a question and formulated a response, thinking aloud as to how they determined the response. Respondents' statements were probed to understand how answers were formulated, to understand the respondents' interpretations of the questions, and to clarify what was or was not included in the response. Respondents were also asked, at times, to describe what technical terms meant to them (e.g., joint custody). At the end of the interview, respondents were debriefed concerning the sensitivity and difficulty of the various questions.

Interviews were conducted at the cognitive laboratory facility of the Center for Survey Methods Research and at the Prince George's County Child Support Office. The interviews were audio-taped and the informants were asked to sign consent forms, permitting review of the tapes by researchers involved in the study. All of the respondents consented to the audio-taping.

Each of the taped interviews was transcribed. In addition, following the interview, the interviewer wrote a detailed summary of the session. These two documents were reviewed by the researcher and used as the primary qualitative information to recommend changes to the questionnaire.

Findings. The findings from the cognitive interviews (reported in Mathiowetz, 1991) suggested the following:

- the new introduction was easy for respondents to answer while also ensuring inclusion of the total universe of eligible persons;

- there did not appear to be any clearly identifiable reason for the high item nonresponse rates for many of the questions, except for the questions which dealt with actual and awarded amounts of child support;

- the main problem with the support questions, which asked the respondent to provide dollar amounts, was that the respondents were forced to provide the answer in terms of an annual amount. Many of the respondents were able to provide weekly or monthly amounts, but had difficulty in annualizing the amount; and

- suggestions for improving the interviewer manual. In light of the lack of cognitive problems with the questions, it was hypothesized that the source of the high item nonresponse rates evident in 1988 and 1990 may be with the interviewer.

Recommendations from this research were incorporated into the April 1992 Child Support Supplement. They were acknowledged as an improvement over the existing questionnaire, and an effective use of the limited amount of time (eight months from planning to analysis and recommendations) and money. However, the sponsor recognized that more research would be necessary to fully address all the data quality problems with the questionnaire, and a longer range program of research is being conducted for the 1994 Child Support Supplement.

B. Medium Length Supplement, Limited Time, Moderate Funds: The Leisure Activity Survey

(Largely extracted from Executive Summary of "A Study of Approaches to Survey Question Evaluation" prepared by Jack Fowler and Tony Roman)

Background. Approximately every five years since 1982, the National Endowment for the Arts (NEA) has sponsored a supplement to the National Crime Survey (NCS) on public participation in the arts, commonly referred to as the Leisure Activity Survey (LAS). The purpose of the LAS is to measure the extent to which American adults attend and participate in various kinds of arts-related performances and activities. The NEA was aware of some problems in the previous version of the survey with respect to respondents' understanding of the kinds of events and activities that they were and were not to report. The NEA was very much interested in having research conducted to address the weaknesses

in the questionnaire, and had budgeted a modest amount of money for a pretest in 1991. The next LAS survey was scheduled to be fielded between January and December 1992.

In late 1990, Census had arranged a Joint Statistical Agreement between Census and the University of Massachusetts' Survey Research Center to do an evaluation study of various pretesting techniques and it was decided that the LAS would be a good vehicle for that evaluation. The LAS schedule required that Demographic Surveys Division (DSD) have the final questionnaire for 1992 no later than mid-June 1991; therefore, there were only five months to plan the research, conduct the pretest, analyze results, propose recommendations and revise the questionnaire. (See Fowler and Roman, 1992, for the detailed report of this project and associated materials.)

Planning. In January 1991, an advisory committee met with the NEA to identify information that needed to be obtained in the 1992 LAS. In late February, the NEA was able to identify specific problems requiring attention. The primary problem was not knowing how broadly or narrowly a respondent is interpreting participation in the arts. There are specific events the NEA wants to exclude and include; however, this is never conveyed to the respondent.

A research schedule was agreed upon by the sponsor, DSD, CSMR and UMASS. The time schedule and staffing and budget resources permitted the research to be conducted in two phases--a

laboratory phase and a field phase. This was decided as the optimal design (for this study) since it would allow time for respondent focus groups and cognitive interviews to be conducted during questionnaire development, prior to the field test. Since focus groups and cognitive interviews are good tools for identifying ambiguities and confusion with survey concepts and problems with comprehension and recall, the results could be used to fine tune the questions during development of the pretest questionnaire. The laboratory research (respondent focus groups and cognitive laboratory interviews) was scheduled for March 1991. Results from the laboratory research would be used for development of an alternative questionnaire that was scheduled to be field tested against the control questionnaire in mid-May 1991. (It was necessary for the alternative questionnaire to be developed by early April so the appropriate interviewer manual, home study and layout of the questionnaires could be ready by mid-May.) The final questionnaire was required to be developed by June 19.

Laboratory Research. Laboratory research was conducted in March 1991 and consisted of two respondent focus groups (eight participants each) and ten cognitive laboratory interviews. (Subjects were recruited by referral by members of the staff of the Center for Survey Research at UMASS.)

In the focus groups, the concepts and terms that were critical to the survey instrument were discussed with an eye

towards identifying ambiguities and confusions. A large proportion of the problematic items pertained only to arts participants, which are a small portion of the general population. Therefore, to ensure that the problematic items were adequately discussed by a relevant population, it was decided to conduct one focus group with known arts participants and the other with persons for whom their level of arts participation was unknown. Focus groups were videotaped and lasted about two hours.

The group discussion went through various kinds of activities covered in the survey and asked people to discuss the type of events or behaviors they thought should or should not be included. Emphasis was placed on identifying ambiguities in the kinds of activities that the NEA wants reported and disagreements among participants on what the various activities included.

Results were produced in two ways. The cognitive psychologist who led the groups wrote a summary of the problems and issues that he identified. Additionally, both he and a research assistant filled out a standardized rating form flagging questions that appeared to be problematic. They rated each question that was discussed in the focus group on whether or not participants seemed to understand the words or ideas in the questions.

After the focus groups were conducted, ten cognitive laboratory interviews were conducted in which a cognitive

psychologist asked probing questions to identify problems with comprehension or recall. Six interviews were conducted with known arts participants and four with persons for whom their arts participation was unknown. As with the focus groups, the laboratory interviews were videotaped.

The cognitive psychologist asked respondents the questions as worded as an initial stimulus. He then used various techniques, including asking respondents to think aloud while they were answering questions and asking follow-up probes to assess the way respondents understood the questions and how they formulated answers. In some cases, he was not able to go through all questions in the two hours allotted. In such cases, he randomly sampled questions in a manner that insured that all questions were tested approximately the same number of times.

The cognitive psychologist conducting the interviews had been informed during development of the interview protocol that interpretation of various subject matter terms used in the questionnaire were of particular interest. Specifically, did respondents have consistent interpretations of terms such as art gallery, operetta, craft fair, dance lessons, etc., or did their interpretations vary? Also of interest was the interpretation of action words used in the questions. For example, did respondents attach a consistent interpretation to such words as listen, watch, visit or go to?

The results of the laboratory research served as input in the development of an alternative questionnaire for the field test. Many questions were reworded, some only slightly and others more extensively. In general the changes can be grouped into the following four categories. They included:

- eliminating unnecessary words to make the questions more compact and easily understood;
- standardizing the inclusions/exclusions in the questions so that the questions refer to the same general universe and, therefore, decrease the amount of potential confusion for the respondent;
- eliminating sentences read after the primary question by building the sentence into the body of the question; and
- eliminating confusing terminology.

It was thought that these general recommendations would produce a set of questions that should be more easily read by interviewers, more easily understood by respondents and less prone to interviewer effects or bias.

Field Test. A split sample field test of the changes in the survey questions was conducted in mid-May. Interviewers administered both versions of the questionnaire. Households were randomly assigned to different questionnaire treatments so that half the respondents answered the control questionnaire, and the other half were interviewed with the test questionnaire that included experimental changes in question wording based on the laboratory research. (It should be noted that the field pretest did not totally simulate the true interviewing environment of the

actual survey since the NCS core questionnaire was not included in the pretest.)

To reflect the procedures of the LAS, two-thirds of the interviews were conducted from a centralized telephone facility and the remaining one-third were conducted through personal visits. (It should be noted that the actual survey instrument used at the centralized telephone facility will be a CATI instrument, while that used during the pretest was a paper questionnaire.)

Washington D.C. was the site of the field test and sample households were selected by the staff in the National Crime Surveys Branch in DSD. Approximately half the sample was selected from a list of patrons of the Kennedy Center, to insure that all questions, including those asked only of arts participants, would be pretested. The remaining 63 households were purposively selected from the same neighborhoods as the arts participants. This served two purposes: it made the "seeding" of the sample (inclusion of known arts participants) transparent to the interviewers, and it cut down on the costs of having to go to different areas.

Five interviewers were selected to conduct the pretest. Prior to interviewing, they completed a home-study that had been prepared by staff in the National Crime Surveys Branch in DSD. Three interviewers administered the questionnaire from the Hagerstown Telephone Center and two interviewers conducted

personal visit interviews. Interviewers were instructed to request permission from respondents to audiotape interviews and tape all interviews for which consent was granted. These tapes were later used for the behavior coding of interviewer/respondent interactions.

Interviewing was conducted between May 13-18. One hundred nine households were interviewed, resulting in 135 interviewed persons. Household members 18 years of age or older were eligible for the survey and self-response was required.

The field pretest interviews were subject to the following evaluation procedures:

Respondent Debriefing - At the end of the interview, all respondents were asked a few extra questions to probe specific areas in which there was concern about respondent comprehension. The purpose of these probes was to find out whether or not respondents were confused or had inconsistent understanding of a few key terms. Specific areas that were probed included attendance at classical music performances, watching classical music performances, visiting historic parks, monuments or neighborhoods, and reading novels, stories or other works of fiction. Data were keyed and tabulated by staff at the UMASS Center for Survey Research.

Behavior Coding - Audiotapes of all completed interviews were coded by coders at the Survey Research Center in Boston. Each question was coded for the following

behaviors: questions were read as worded; respondents asked for clarification; respondents initially gave an inadequate answer; and respondents interrupted the reading of the question. Coding was completed in time for the results to be incorporated into the debriefing discussion with the interviewers.

Interviewer Debriefing - A group setting interviewer debriefing was conducted on May 23, after the pretest interviews had been completed. Interviewers reported on various problems and issues they found during administration of the field test. After staff from Field led the interviewers through a question-by-question evaluation, Tony Roman from UMASS presented the results obtained from behavior coding analysis of the taped interviews.

Interviewer Rating Forms - Interviewers were provided with interviewer rating forms prior to the group debriefing. These forms provided a systematic method for the interviewers to flag questions they thought were problematic with respect to wording, respondent comprehension or the task respondents were given.

Response Distributions - The actual responses from the pretest interviews were coded and the distributions of responses were available for use in the question evaluation process.

Results. Results from the laboratory research (respondent focus groups and cognitive laboratory interviews) indicate that these methods are effective techniques for gaining information about questions. Both led to information that would probably have been missed in a conventional pretest and both methods proved helpful in identifying problems with comprehension and with answering questions accurately. While the focus groups were useful in identifying ambiguities in terms and concepts, the laboratory interviews were best to gain insight into difficulties respondents had with the particular response task they were given. In this research, the two procedures were found to be highly complementary.

Results indicated that many question problems identified in the field pretest had not been identified in earlier testing. This demonstrates that laboratory methods alone are not sufficient to test a questionnaire. While the focus groups and laboratory interviews were helpful in evaluating terms, concepts, and question wording under controlled laboratory conditions, it was necessary to determine how well the questions would work in a realistic setting with representative interviewers and respondents through a field pretest.

Table 1 compares problems identified from behavior coding, rating forms and interviewer debriefings. The pre-field techniques are not included in this comparison since data for all questionnaire items were not always obtained from all

participants in the focus groups and cognitive interviews. In addition, the questions were revised between the pre-field research and the field test, prohibiting a direct comparison. Additionally, the respondent debriefing administered during the field test is not included in this comparison since the respondent debriefing applied to a very small subset of the questions in the alternative questionnaire. As the table shows, there was considerable overlap in the questions flagged by these methods. All the items flagged in the rating forms were also mentioned in the interviewer debriefing. Also, interviewers mentioned many more problems in their rating forms and in the debriefing than showed up as problems from the behavior coding. One type of problem that was uniquely reported in the debriefing was the way questions fit together. Issues relating to question flow or transitions between sections surfaced here. The behavior coding was most helpful in identifying questions that were consistently misread by interviewers or interrupted by respondents. The respondent debriefing questions helped identify problems not found by the other methods, most notably that persons were counting nonfiction books they read as "novels" when asked about novels read. Response distribution data compared distributions of the old and new questions to get some idea of the effect of new wording on the resulting data. However, due to small sample sizes there was limited confidence that the results were reliable.

TABLE 1: Number of Potentially Problematic Questions Discovered During LAS Pretest by Pretesting Method Used

	With no Other Method	Also in Behavior Coding	Also on Rating Forms	Also in Debrief-ing	In Both Other Methods	TOTAL
Behavior Coding	5	-	0	7	18	30
Rating Forms	0	0	-	34	18	52
Debrief-ing	20	7	34	-	18	79

SOURCE: FOWLER AND ROMAN, 1992

The pretest research led to several changes in the questionnaire, some of which are displayed in Table 2 for illustrative purposes. The most significant revision was to clearly state within the question what legitimately should or should not be included within the activity being asked about. Additional questions were included to serve as screeners for later questions so the most appropriate universe of persons were asked the questions of interest. In several questions, activities being asked about were more clearly defined within the question. The NEA, DSD and CSMR staff agreed that overall the changes seemed to make the questionnaire flow more smoothly and the changes did more to guarantee that respondents were referencing the same response universe.

TABLE 2. Examples of Changes to LAS Questionnaire Based on Pretesting

Original Question Wording	Revised Question Wording
(During the LAST 12 MONTHS,) Did you go to a live dance performance other than ballet? This includes modern, folk, tap, or other dance.	(With the exception of elementary or high school performances), did you go to a live dance performance other than ballet, such as modern, folk, or tap during the LAST 12 MONTHS?
(During the LAST 12 MONTHS), Did you read any - Novels or short stories? Plays? Poetry?	With the exception of books required for work or school, did you read any books during the LAST 12 MONTHS? (If yes then) (During the LAST 12 MONTHS,) Did you read any- Plays? Poetry? Novels or short stories?
(During the LAST 12 MONTHS,) Did you visit an ART or craft fair or festival?	(During the LAST 12 MONTHS,) Did you visit an ART fair or festival, or a craft fair or festival?

C. Extensive questionnaire, time for several phases of pretesting: CPS Tobacco Use Supplement

Background. This study was done on a supplement to the Current Population Survey that concerns tobacco use, sponsored by the National Cancer Institute (NCI). The supplement was scheduled to be administered in September 1992 and in January and May 1993.

The questionnaire, which was designed for self-response, contained questions about the respondents' use of tobacco products, their perceptions of smoking policies in their workplace, and their attitudes about smoking policies and practices. It represented a new questionnaire, but it did contain some questions that are also asked on the National Health Interview Survey. As researchers, we were somewhat constrained in our ability to make changes to the parallel questions due to the sponsor's desire for comparative data across the two surveys.

Planning work began in the fall of 1990 with the negotiation of the interagency agreement. The final pretest plan included both cognitive interviews conducted in a laboratory environment and a field pretest that yielded results on the survey itself, the behavior of interviewers and respondents, and feedback from the field staff. Cognitive interviews were used to gain in-depth insights about respondent problems with the questionnaire. This was particularly important because there was a large battery of attitudinal items that were newly-developed and had never been tested. This methodology would provide large amounts of information that could be used to revise the questions. However, these interviews would not replicate the field situation, and once the revised questionnaire was developed, a field test was necessary to evaluate it under actual conditions.

Preparatory work for cognitive interviewing was done in February and March 1991; 21 interviews were conducted over a two-

week period in April and one staff member conducted 6 additional interviews in Spanish in the fall. Spanish interviews were conducted because the sponsor requested that a standardized Spanish translation of the questionnaire be used in the field (rather than the traditional practice of ad hoc translation by relatives or neighbors of the respondent). The sponsor provided the translation.

The field pretest and the interviewer debriefing took place in November 1991; the behavior coding was completed in December.

Cognitive Interviews. During the cognitive interviewing phase, five staff members from different divisions in the Census Bureau participated (DSD, FLD, CSMR). All were familiar with standard interviewing practice, and received additional training to learn to conduct cognitive interviews. They met as a group to review the questionnaire in detail, deciding on goals for the interviews and discussing appropriate questioning techniques and probes to elicit the desired information. They developed a protocol that included follow-up probes to determine how respondents interpreted specific reference periods (e.g., "in the past year"), how they interpreted specific terms (e.g., "fairly regularly"), and how confident they were in their answers (e.g., how old they were when they first started smoking).

The questionnaire included separate sections for current smokers, former smokers, and persons who never smoked. All three types of respondents were recruited to test all the questions,

and also to test whether the attitudinal questions (asked of everyone) were problematic for persons in a particular smoking status. In order to recruit respondents from the local community for the English language interviews, brightly colored posters were placed in a nearby shopping mall, a fitness center, a medical clinic, a library, and non-Census offices in the building where the laboratory is located. An ad was also placed in the Capital Flyer, which is a newspaper distributed at nearby Andrews Air Force Base. It reaches both active duty and retired military personnel as well as civilians who work on the base. This proved to be the best source for locating respondents. More than three-quarters of the volunteers came from this source. Volunteers were asked to call for an appointment. Two support staff members screened the callers on smoking status and scheduled the interviews. Staff from the National Cancer Institute took care of recruiting for the Spanish language interviews that were to take place on the campus at the National Institutes of Health (NIH) in Bethesda, MD.

We experienced very little difficulty with "no shows" for interviews conducted at CSMR's laboratory, but we did find it necessary to do some re-scheduling because a few respondents could not keep their original appointments. We had more problems with the few respondents scheduled to be interviewed in Spanish. Of the original six persons scheduled, only three

came at the appointed time; three additional respondents were recruited on an "emergency" basis through a contact in the office at NIH.

Even though we screened callers only on smoking status, we achieved an age range of respondents from the teens through the sixties. However, we interviewed over 70 percent females and 90 percent whites in the English- language interviews. In the Spanish-language version, one-third of the respondents were male and the ages ranged from the twenties through the fifties.

Interviewing. Each interviewer greeted the respondent, gave a brief explanation of what would be required, demonstrated the "think aloud" technique for answering, assured confidentiality, and secured permission for taping. (All sessions were audiotaped and the interviews conducted at CSMR's laboratory were also videotaped.) Once the taping equipment was operating, the interview proceeded. Interviewers recorded item responses on a paper questionnaire and relied on the tape for a record of the interview and follow-up probes.

Interviews were scheduled in one-hour timeslots (the average interview length of the final questionnaire was expected to be six minutes, with variations among current smokers, former smokers, and non-smokers). In most cases, this time proved to be adequate: actual interview times ranged from 15 minutes to nearly an hour.

Interviewers generally followed the protocol, using the standard probes that had been previously developed. In addition, the specific content of respondents' answers also required that unique probes be used to obtain a full understanding of their answers and the problems they had.

Documentation, analysis, and presentation of results. Each interviewer used the audiotape and the questionnaire to prepare a detailed item-by-item summary of the interview. (Videotapes were available for viewing in place of "live" observation.) When all interviewing summaries were completed, the group of researchers met for a short debriefing and agreed on a method to prepare the final report. Each interviewer took responsibility for writing an item-by-item review of the problems encountered in a section of the questionnaire summarized across all interviews. After all participants had reviewed these summaries, we met again to develop recommendations for modifications to the questionnaire.

The final report (DeMaio, et al, 1991) included the item-by-item review of each section of the questionnaire, our recommendations for modifications, and the original interview summaries as an attachment. The report of the Spanish interviews (Glover, 1991) was organized the same way.

Cognitive interviewing revealed the following types of weaknesses in our questionnaire:

- awkward or ambiguous question wording;

- variability in respondents' interpretations of question meaning;
- inadequate or unnecessary response categories;
- the need for screener questions and/or additional skip patterns; and
- problems with flow and question order.

In carrying out this study in our response research laboratory, we also learned that we need to expand the methods and sources we used to recruit respondents, and that respondents did not object to either the audio- or video-taping.

After the report on the results of the cognitive interviewing was distributed, questionnaire revisions were negotiated with the sponsor. A meeting with staff from NCI, DSD, and CSMR was held to review the recommendations contained in the report. These recommendations led to some clarifications in question wording (e.g., "trying to quit"), some minor changes including rearranging phrases within a question, revising one entire section (on smoking policies in the respondent's workplace), and omitting some of the problematic attitudinal questions.

Field Test. A field test of the revised questionnaire was conducted in the fall of 1991 to see how it worked under actual field conditions. The sample was selected from outgoing rotations of CPS from August, September, and October 1991 for the Washington, D.C.-MD-VA metropolitan statistical area. The

sponsor's original request was for interviews with 300 persons, broken down into specific numbers of cases by smoking status, age, race/ethnicity and education. This was not feasible, so instead a larger sample (300 households or approximately 600 persons) was used, with no previous screening for respondent characteristics. Based upon known national distributions by smoking status, this was thought to provide adequate numbers of interviews with current and former smokers. The selection of PSUs for interview was also made to maximize the diversity of the sample by race/ethnicity.

Ten interviewers and one supervisor from the Charlotte region participated in the pretest. Interviewers completed a home study prior to the classroom training session on October 30, 1991.

Sample households were sent an advance letter before interviewing began. From November 1-14, interviewers completed 450 person interviews in 260 households, with 47 Type A noninterviews. The response rate was 84.7 percent.
Interviewers were instructed to tape record interviews after securing the respondent's permission to tape. They were provided with tape recorders and a device to allow taping of telephone interviews. The tapes were collected to facilitate coding of the interviewer and respondent behaviors.
They were also asked to keep a "master" copy of the questionnaire on which to record notes of problems they experienced with the

questions. This "master" was used to help their recall in the debriefing scheduled to follow the interview period.

Interviewers first asked the CPS labor force items of the household respondent; all household members age 15 and over were eligible for the self-response supplement. Proxy interviews for the supplement were accepted only after two callback attempts. Interviews were designated for either telephone or personal visit to achieve a 75/25 split.

The field test interviews were subject to the following evaluation procedures:

Interviewer Debriefing - Staff from the Census Bureau and from NCI attended the interviewer debriefing session held on November 15, 1991. Interviewers were asked to focus on the questionnaire section by section and to offer their observations of the interview process. The session was tape recorded for later documentation.

Field representatives made the following observations during the debriefing:

- they had problems with the placement of skip instructions in some questions;
- some questions were too long, or response lists were too long: they sensed that respondents had not remembered enough of what was being asked to answer the question accurately;
- they experienced problems with question order and flow in parts of the questionnaire;

- the questions sometimes employed terms that were unclear to respondents; and

- they had difficulty rewording questions for administering to proxy respondents.

Behavior Coding - The sample of 150 taped interviews was stratified by smoking status, mode of interview, type of respondent (self vs. proxy), and interviewer. Six professional staff members of the CSMR, DSD, and FLD coded the 150 sampled interviews. Staff in CSMR carried out the sampling, arranged assignment materials for the coders, and keyed the codes in a dBase4 datafile that was later converted to a SAS file for analysis.

Each of the coders completed 25 interviews. Seven were coded shortly after the start of the field period so that some preliminary results could be made available for discussion at the field staff debriefing. The balance of the work was completed after the debriefing. Additional interviews were coded by all coders to supply reliability measures for their work at the beginning and the end of the coding operation. The reliability measures were not actually calculated until the data were analyzed. In hindsight, this was not optimal timing, because the results showed that one coder coded the cases very differently than the others. If the analysis had been done earlier, these differences would have been observed and further

training could have been given to assure that all coders were coding the same way.

The behavior coding provided the following information (Mathiowetz, et al, 1991):

- asking respondents to recall exact ages and amounts of elapsed time resulted in high rates of qualified or inadequate answers, indicating that these types of questions were problematic;

- the format and flow of the questions needed to be improved, as indicated by the high rates of major changes to question wording on the part of interviewers;

- at times the intent of the question was not clear and respondents requested clarification;

- response categories that were not all-inclusive or mutually exclusive caused high rates of qualified or inadequate answers;

- questions that were too long caused problems both for interviewers (major changes in wording) and respondents (inadequate answers); and

- interviewers introduced major wording changes when posing questions to proxy respondents.

Item Nonresponse and Response Distribution Analysis - Data from the questionnaires were keyed and a datafile was created. Item nonresponse and response distribution tallies were also produced for analysis.

Item nonresponse rates were calculated for the entire sample and broken down by mode of interview (telephone, personal visit) and respondent status (self,

proxy). Tallies of response distributions were also generated. The results of reviewing these data showed that:

- there were a number of items with high nonresponse rates, and these items were also identified as problematic in the behavior coding results;
- item nonresponse was more of a problem for the personal visit cases than for the telephone interviews;
- there were a number of items that suffered from high "don't know" rates, indicating that respondents couldn't answer the questions. Although many of these items were revealed as problematic through other evaluation sources, this was not true in all cases; and
- rates of "don't know" responses were higher for proxy than for self-response interviews.

Based on the results of the field test, questionnaire changes were made to address some of the problems uncovered (e.g., proxy problems, format and flow problems, problems with response categories and confusing intent). Other problems (e.g., recall of ages and elapsed times) were not dealt with because those questions were also contained on the HNIS questionnaire and comparability of data was viewed as having a higher priority.

The work on the Tobacco Use Supplement for the CPS was among the first of our studies employing several phases of pretesting to improve survey questionnaires. Cognitive interviewing proved to be a valuable first stage in the pretest process. It allowed us to inexpensively identify and resolve problems in the

questionnaire before the more costly field test. The findings from the behavior coding, the interviewer debriefing, and the analysis of responses from the field test all further improved the final questionnaire. The problem questions identified through behavior coding generally overlapped with those identified through the analysis of responses. Due to extensive written comments by the behavior coders, this method was also useful in identifying the sources of the problems. The interviewers in the debriefing session provided additional guidance in identifying the causes of some of their problems.

VI. Summary and Conclusions

In planning a program of pretest activities, time and cost will play a large role in determining what is feasible. To the extent that alternative scenarios can be entertained, we recommend that both pre-field and field techniques be undertaken. As noted in our discussion of the Leisure Activities Survey (Case Study #2), problems discovered in field tests had not always been identified during pre-field testing. This suggests that laboratory methods alone are not sufficient to test a questionnaire.

In terms of pre-field techniques, the choice is between focus groups and cognitive interviews. For continuing surveys that have a pre-existing questionnaire, cognitive interviews should be used to provide detailed insights into problems with

the questionnaire whenever time permits. In our minds, these are more useful than focus groups at the same stage of questionnaire development because they mimic the question-response process. The advantages of focus groups come at an earlier stage, when information about the concepts, terminology, and sequence of topics are still in flux for one-time or new surveys. At this point, focus groups provide researchers drafting the questionnaire with useful information about how to structure the questionnaire and word the questions.

In terms of field techniques, all the methods outlined in Section III can be incorporated into a field test with minimal cost, and we encourage the use of as many as possible.

There are some situations in which it is not feasible to use some of the methods. For example, split panel tests are appropriate only when a control questionnaire exists for testing against a new one, or when theoretical or practical considerations dictate the development of alternative questionnaires. Also, respondent debriefing is limited by the extent to which hypotheses are available about potential respondent problems suitable for probing. (One source of such information is cognitive interviews.)

The remaining methods--behavior coding, interviewer debriefing, and analysis of responses--each contribute to the evaluation of the field test in different ways. Their contributions also vary depending on how many questionnaires are

being tested. Interviewer debriefing provides information (qualitative or quantitative, depending on how it is implemented) from the point of view of the interviewers. The advantage is that interviewers are part of the data collection process and they build up a wealth of information about respondents and questionnaires during their work. The disadvantage is that there may be value judgments or recall biases that affect the information they provide. Behavior coding is an objective method of monitoring the interaction between the respondent and the interviewer to obtain quantitative data about the nature and frequency of problems. It does not always provide guidance about the causes of problems, although comments from the coders are useful in that regard. Analysis of responses for item nonresponse rates, frequency distributions, and "don't know" rates also provides evidence of problems but no information about causes. And when respondent debriefing is introduced, it can provide information about specific types and causes of problems. Melding the objective with the subjective methods, and the respondent-centered with the interviewer-centered methods provides an evaluation of broad scope. The methods, with their associated strengths and weaknesses outlined in Section III, complement each other with respect to problem identification and problem source. Differences in sample size, questionnaire design, and mode of data collection may affect the relative importance of the various techniques for evaluating survey

questions. By incorporating as many techniques as appropriate in planning a field test, the maximum benefit can be derived from the evaluation.

REFERENCES

Abraham, S., "The Use of Laboratory Techniques in the Questionnaire Pretest Process: The Chicago Urban Survey Experience," paper presented at the Annual Conference of the American Association for Public Opinion Research, 1989.

Belson, W., *The Design and Understanding of Survey Questions*, Aldershot, England: Gower, 1981.

Biemer, P., R. Groves, L. Lyberg, N. Mathiowetz, and S. Sudman (eds.), *Measurement Errors in Surveys*, New York: John Wiley and Sons, 1991.

Bischoping, K., "An Evaluation of Interviewer Debriefing in Survey Pretests," Chapter 2 in C. Cannell, L. Oksenberg, F. Fowler, G. Kalton, and K. Bischoping, *New Techniques for Pretesting Survey Questions.* Final Report for Grant Number HS 05616 from the National Center for Health Services Research and Health Care Technology Assessment. Ann Arbor, MI: Survey Research Center, University of Michigan, 1989.

Bradburn, N., S. Sudman and Associates, *Improving Interview Method and Questionnaire Design*, San Francisco: Jossey-Bass Publishers, 1979.

Campanelli, P., E. Martin, and K. Creighton, "Respondents' Understanding of Labor Force Concepts: Insights from Debriefing Studies," *Proceedings of the Fifth Annual Research Conference*, U. S. Bureau of the Census, 1989, pp. 361-374.

Campanelli, P., E. Martin, and J. Rothgeb, "The Use of Respondent and Interviewer Debriefing Studies as a Way to Study Response Error in Survey Data," *The Statistician*, Vol. 40, 1991, pp. 253-264.

Cannell, C., S. Lawson, and D. Hausser, *A Technique for Evaluating Interviewer Performance*, Ann Arbor, MI: Survey Research Center, The University of Michigan, 1975.

Converse, J., and S. Presser, *Survey Questions: Handcrafting the Standardized Questionnaire*. Sage University Paper Series, Number 63, Beverly Hills: Sage Publications, 1986.

Cook, T., and D. Campbell, *Quasi-Experimentation, Design and Analysis Issues for Field Settings*. Boston: Houghton Mifflin Co., 1979.

DeMaio, T. (ed.), *Approaches to Developing Questionnaires,* Statistical Policy
Working Paper 10, Washington, DC: Office of Management and Budget, 1983.

DeMaio, T., S. Ciochetto, L. Sewell, M. Beach, and T. Glover, "Report of Results of Cognitive Interviewing for the CPS Tobacco Use Supplement for the ASSIST Evaluation," internal Census Bureau report, June 26, 1991.

Esposito, J., and J. Hess, "The Use of Interviewer Debriefings to Identify Problematic Questions on Alternative Questionnaires," paper presented at the Annual Meeting of the American Association for Public Opinion Research, 1992.

Esposito, J., J. Rothgeb, and P. Campanelli, "Instructions for Using the CPS CATI/RDD Monitoring Form," Unpublished document, 1991.

Fleiss, J., *Statistical Methods for Rates and Proportions*, New York: John Wiley and Sons, 1981.

Forsyth B., and J. Lessler, "Cognitive Laboratory Methods: A Taxonomy" in P. Biemer, R. Groves, L. Lyberg, N. Mathiowetz, and S. Sudman (eds.), *Measurement Errors in Surveys*, New York: John Wiley and Sons, 1991.

Fowler, F., "Reducing Interviewer-Related Error Through Interviewer Training, Supervision, and Other Means," in P. Biemer, R. Groves, L. Lyberg, N. Mathiowetz, and S. Sudman (eds.), *Measurement Errors in Surveys*, New York: John Wiley and Sons, 1991.

Fowler, F., "Evaluation of Special Training and Debriefing Procedures for Pretest Interviews," in C. Cannell, L. Oksenberg, F. Fowler, G. Kalton, and K. Bischoping, *New Techniques for Pretesting Survey Questions.* Final Report for Grant Number HS 05616 from the National Center for Health Services Research and Health Care Technology Assessment. Ann Arbor, MI: Survey Research Center, University of Michigan, 1989.

Fowler, F., *Survey Research Methods*. Applied Social Research Methods Series, Volume 1, Beverly Hills: Sage Publications, 1988.

Fowler, F., and T. Mangione, *Standardized Survey Interviewing: Minimizing Interviewer-Related Error*. Applied Social Research

Methods Series, Volume 18, Beverly Hills: Sage Publications, 1990.

Fowler, F., and A. Roman, "A Study of Approaches to Survey Question Evaluation," Center for Survey Research, University of Massachusetts, 1992.

Glover, T., "Report of Results of Spanish Cognitive Interviewing for the CPS Tobacco Use Supplement for the ASSIST Evaluation," internal Census Bureau report, October 18, 1991.

Groves, R., M. Berry, and N. Mathiowetz, "The Process of Interviewer Variabililty: Evidence from Telephone Surveys," *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 1980. pp. 519-524.

Hubbard, M., J. Lessler, K. Graetz, and B. Forsyth, "Improving the Comprehension of Reference Periods," *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 1990, pp. 474-479.

Jabine, T., M. Straf, J. Tanur, and R. Tourangeau (eds.), *Cognitive Aspects of Survey Methodology: Building a Bridge Between Disciplines*, Washington, DC: National Academy Press, 1984.

Lessler, J., R. Tourangeau and W. Salter, "Questionnaire Design in the Cognitive Research Laboratory," *Vital and Health Statistics*, Series 6, No. 1, 1989, Washington, DC: National Center for Health Statistics.

Loftus, E., S. Feinberg, and J. Tanur, "Cognitive Psychology Meets the National Survey," *American Psychologist*, February 1985, pp. 175-180.

Martin, E., and A. Polivka, "The Effect of Questionnaire Redesign on Conceptual Problems in the Current Population Survey," paper presented at the Annual Meeting of the American Statistical Association, 1992.

Mathiowetz, N., "Report on Cognitive Research on the Child Support Enforcement Questionnaire--April Supplement to the Current Population Survey," internal Census Bureau report, June 10, 1991.

Mathiowetz, N., S. Ciochetto, T. DeMaio, and L. Sewell, "Report of Results of Behavior Coding and Interviewer Debriefing," internal Census Bureau report, February 5, 1992.

Mathiowetz, N., and C. Cannell, "Coding Interviewer Behavior as a Method of Evaluating Performance," *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 1980. pp. 525-530.

Morgan, D., *Focus Groups as Qualitative Research*, Qualitative Research Methods, Vol. 16, Newbury Park: Sage Publications, 1988.

Morton-Williams, J. and W. Sykes, "The Use of Interaction Coding and Follow-up Interviews to Investigate Comprehension of Survey Questions," *Journal of the Market Research Society*, Vol. 26, 1984, pp. 109-127.

Nelson, D., "Informal Testing as a Means of Questionnaire Development," *Journal of Official Statistics*, Vol. 1, 1985, pp. 179-188.

Oksenberg, L., C. Cannell, and G. Kalton, "New Strategies for Pretesting Survey Questions," *Journal of Official Statistics*, Vol 7, No. 3, 1991, pp. 349-365.

Payne, S., *The Art of Asking Questions*, Princeton, NJ: Princeton University Press, 1951.

Presser, S., and J. Blair, "Survey Pretesting: Do Different Methods Produce Different Results?", unpublished manuscript, April 1993.

Rothgeb, R., and J. Hess, "The Role of Response Distribution and Item Nonresponse Analysis in Evaluating Alternative Question Wordings During the Redesign of the Current Population Survey," paper presented at the Annual Meeting of the American Association of Public Opinion Research, 1992.

Schuman, H., and S. Presser, *Questions and Answers in Attitude Surveys: Experiments on Question Form, Wording, and Context*, New York: Academic Press, 1981.

Sudman, S., and N. Bradburn, *Asking Questions: A Practical Guide to Questionnaire Design*, San Francisco: Jossey-Bass Publishers, 1982.

Sudman, S., and N. Bradburn, *Response Effects in Surveys: A Review and Synthesis*, Chicago: Aldine, 1974.

Turner, C., and E. Martin (eds.), *Surveying Subjective Phenomena*, Vols. 1 and 2, New York: Russell Sage, 1984.

Appendix A

Description of Supplementary Cognitive Laboratory Techniques

Memory cue tasks are a strategy for stimulating recall. The terms refers to the technique of providing the respondent with detailed memory cues to elicit either more details about an event or the enumeration of more "events" that may otherwise be forgotten. In memory cue tasks, respondents are initially asked a "free recall" question (e.g., "During the past 12 months, what was your total household income?") and then are provided cues (e.g., "What about income from interest or dividends?") to aid their memory search. Memory cue tasks have been used to reduce recall error by providing cues during the survey interview.

Free sort tasks are used to understand the organization of memory and how respondents might conceptualize various topics targeted for a questionnaire. Thus survey practitioners may better understand how to organize a sequence of questions within a questionnaire or bridge between topics.

> In free-sort classification, respondents are asked to sort a set of cards into groups. The cards list the questionnaire items of interest. No instructions are given to the respondent concerning the criteria to be used in sorting the cards. For example, respondents might be given a set of cards containing types of food items and asked to sort them into piles without any criteria.
>
> In contrast, in dimensional-sort classification the respondent is given a set of criteria by which to sort the cards--for example, least to most sensitive, least to most difficult to recall. Both free and dimensional sort classification provide insight as to how respondents think about the topics of interest, and the results can be used to structure the sequence of topics or questions within a questionnaire. Respondents in the previous example might be asked to sort the cards containing food items according to similarity. The results might be used to identify redundancies and eliminate food items from the questionnaire.

Vignette classifications are similar to sort classifications, but they require that the respondent sort through more detailed information. Respondents either read through or listen to short descriptions of situations ("vignettes") and then select a classification that best describes the situation. For example,

vignette classification has been used to determine how both interviewers and respondents classified individuals according to labor force participation (Campanelli, et al, 1991). In this case, vignettes described individuals in different work situations (e.g., a student who painted a house last week and received $20 from his father); interviewers and respondents then classified the individuals as working or not working.

Response latency measurement involves measuring the amount of time between the reading of a question and a response. (Such measurement requires special equipment or a self-administered computer-assisted interview which can record the elapsed time between the question appearing on the screen and the entry of a response). The measurement, based on cognitive retrieval theory, provides an indication as to how difficult the recall task is for the respondent. It should be noted that other factors besides task difficulty (e.g., respondent's rate of speech or interviewer's typing ability) may affect response time.

Appendix B

Conducting "Think Aloud" Interviews:
Planning and Carrying Out Individual Sessions

PLANNING ACTIVITIES

Topic/Task

- Specify the topics or tasks to be covered in the session. For example the session might be devoted to gaining a general understanding of how respondents think about a particular topic, the language they use, and the sophistication of their thinking in preparation for developing a questionnaire. If one is evaluating an existing questionnaire, one may be interested in investigating understanding, in observing the recall process, in determining how respondents make judgements based upon their recall and their interpretation of question, and so on. Prior to the session, one needs to determine the topics that are to be covered and outline the issues that are to be investigated in the session. If a survey addresses sensitive topics, a crucial issue is the respondent's willingness to discuss sensitive issues.

- List the goals of the session. Generally, an entire questionnaire cannot be covered in a single session; so it is necessary to establish specific goals for each session. Initially, one might look at particular topics in isolation. As decisions are made concerning certain topics, one might wish to do some work on the entire questionnaire. However, given the time that it takes to collect a think aloud interview, it will generally not be possible to cover an entire questionnaire.

- An interviewer's guide should be developed, including specific probes that will be used to query the respondent on his or her recall process, understanding of terms, and, possibly, reluctance to answer certain types of questions. For example, you might want to determine whether respondents have thought about a particular topic. You might want to investigate their understanding of particular terms, their reactions to survey procedures, and so on. Because of this, the list of goals for a series of sessions should be specific.

Respondents

- Specify the types of respondents to be included in the individual interviews. Include the age, sex, race, SES, and

other characteristics that you think might influence the response. Specify number of people of each type to be interviewed.

- Recruit the respondents and inform them of the topics to be covered, the meeting place, the time required, and their payments.

Task Materials

- Develop questionnaires, handouts, and other materials that you will need to use in the session. Make copies. You will need an introduction to the respondent that explains what the goals of the session are, who the sponsor is, and describes the voluntary nature of the session. You will also need an informed consent that explains that everything is voluntary. At the end of the session respondents should be given the opportunity to refuse for anyone else to listen to his or her tape.

- Many respondents will benefit from a demonstration of what you want them to do. One way to do this is to show a brief tape of someone thinking out loud as they respond to questions; another is to demonstrate yourself by using a question unrelated to the topic of interest.

Interviewer's Guide

- Develop an interviewer's guide for use in the interview. This should contain a list of probes and questions that are directed at the particular topics under investigation. It will be necessary to "go with the flow" of the interview. Some people will get the idea of what you want right away; others will need more specific probing and encouragement.

- Establish times to be devoted to each component of the session.

Conducting the Interviews

- Arrange the room so that the information can be recorded. It will be important that the session be private if sensitive topics are covered.

- Conduct the session using the interviewer's guide. It is best to focus on the respondent during the session and not try to write down a lot of what the respondent says. Depend on the tape to do the documentation. The key thing to look out for is not leading the respondents while encouraging

them to report their thought processes and feelings as they respond.

- If any sensitive topics are covered in the session, the interviewer should be alert to the feelings of the respondents and make sure their privacy is protected.

SUMMARY AND ANALYSIS

Review of the Results

- Summarize the results of the session. The tapes can be transcribed if a detailed analysis is needed. Otherwise the tape can be reviewed in order to produce the summary. The easiest time to write the summary is immediately after the session because your own memory is more clear. It takes a secretary about 4 to 6 hours to transcribe an hour session. It will take the researcher about 2 hours to review it and write up notes.

Evaluate the Session

- Evaluate the interview in terms of the goals and any problems that arose during the interview. Modify the procedures as necessary. Listening to the tapes is a good way to determine what types of probes enhance the collection of response protocols.

Make a File of the Results

- File the materials, the recordings, the lists of respondents, and the reports. Make sure the recordings and the lists of respondents are in locked drawers.

Confidentiality

- Never discuss the results by name. Also, all discussion of results should be in professional settings. You will hear things that make for good "party talk;" however, you can't mention them. The data are as confidential as questionnaire data.

Appendix C

Expanded Interview Behavior Codes

Category	Code	Used When the Interviewer....
1-Asks question on the as printed	11	reads question exactly as printed questionnaire
	12	reads question making minor modifications of the printed version, but does not alter frame of reference
2-Asks question printed, but incorrectly	21	reads main stem of question as modifies or incorrectly reads any response categories in the question (does not apply therefore to open questions, since they do not have response categories)
	22	either significantly alters main body or stem of question while reading it, or reads only part of it
	23	does not read question, but instead makes a statement about the response he anticipates
	27	asks a question which should have been skipped
3-Probes or which is clarifies non-directively	31	makes up in own words a probe (query) non-directive
	32	repeats printed question or part of it correctly
	34	repeats respondent's response, or part if it, correctly

	35	confirms a frame of reference for respondent correctly and in a non-directive manner
4-Probes or limiting or clarifies either the response	41	makes up a probe which is directive, changing the frame of reference of question or the potential directively
	42	either repeats question and/or response choices incorrectly or gives incorrect summary of respondent's response
	43	gives an introduction which is directive
	45	either interprets question by rewording it or confirms a frame of reference incorrectly
5-Other for clarification behavior	51	helps respondent to understand his role, appropriate example by task-oriented
	58	exhibits other acceptable behavior, such as volunteering general feedback
6-Other inappropriate behavior	62	interrupts respondent
	63	gives personal opinion or evaluation
	67	records response incorrectly or incompletely on questionnaire
	68	exhibits other unacceptable behavior
7-Non-recorded	71	omits question correctly (due to skip pattern)
	72	omits question incorrectly
	73	writes in inferred or previously obtained answer
	75	fails to probe after inadequate answer
	78	missing data, no sound on tape

8-Pace and voice inflection	81	reads question more slowly than 2 words/sec.
	82	reads question at 2 words/sec.
	83	reads question more quickly than 2 words/sec.
	84	conducts entire interview too slowly
	85	conducts entire interview at right pace
	86	conducts entire interview too quickly
	87	reads questions in a wooden, expressionless manner
	88	reads questions with a rising inflection at the end
	89	reads questions with voice dropped, so that they sound like a statement
9-Background of study	91	mentions own name
	92	mentions sponsorship
	93	mentions anonymity
	94	mentions respondent selection procedures
	95	mentions purpose of study

Appendix D

ID#

Behavior Coding Form

_____

Question Reading Respondent Behaviors

Question Number	E	S	M	V	WV	IN	CL	AA	QA	IA	DK	RE	Notes
2	X									X			
							X						"fairly regularly" - WM
								X					

IN: Interrupt, CL: Clarify, AA: Adequate Answer, QA: Qualified Answer, IA: Inadequate Answer

Appendix E:

Sample Behavior Coding Analysis

Table 1. Percent Behavior Codes for Smoking Prevalence/Screener Questions[a]

	Question Number					
	Q1	Q2	Q3	Q4	Q5a	Q5b
Interviewer Behavior[b]						
Question Read as worded	84	83	72	73	16	50
Slight change in wording	15	7	14	22	11	0
Major change in wording	1	10	12	4	74	50
Answer verified correctly	1	0	1	0	0	0
Answer verified incorrectly or not at all	0	0	0	0	0	0
N	150	100	98	45	19	8
Respondent Behavior[b,c]						
Adequate answer	97	73	98	93	80	75
Qualified answer	1	19	1	9	20	0
Inadequate answer	3	14	1	16	20	25
Interruption	1	1	12	0	0	0
Clarification	2	1	0	0	20	0
Don't know	1	0	0	0	0	0
Refusal	0	0	0	0	0	0

[a] Includes both personal and telephone interviews as well as both self and proxy responses.

[b] Percentages which not add to 100 percent due to rounding.

[c] Percentages may be greater than 100 percent if more than one respondent behavior was recorded for the item.

N[d]	146	88	85	45	5	4

[d] The N for respondent behaviors reflects the omission of responses for questions with major changes in wording or wrong verifications.

Appendix F

INTERVIEWER RATING FORM

Use the following code for each potential problem:

A No evidence of problem
B Possible problem
C Definite problem

COLUMN 1 Should be used for potential problems due to having trouble reading the question as written.

COLUMN 2 Should be used for potential problems due to respondents not understanding words or ideas in the questions.

COLUMN 3 Should be used for potential problems due to respondents having trouble providing answer to question.

	Column 1	Column 2	Column 3		
Question Number	Hard to Read	R has problem understanding	R has trouble providing answer	Other problems	Comments

APPENDIX G

PROTOCOL FOR PRETESTING DEMOGRAPHIC SURVEYS AT THE CENSUS BUREAU

Report of the Pretesting Committee
Theresa DeMaio, Chair
Nancy Mathiowetz
Jennifer Rothgeb
Mary Ellen Beach
Sharon Durant

with contributions from Floyd J. Fowler and Anthony M. Roman,
Survey Research Center, University of Massachusetts at Boston

June xx, 1993

TABLE OF CONTENTS