THE USE OF COGNITIVE LABORATORY INTERVIEWS FOR ESTIMATING PRODUCTION SURVEY COSTS AND RESPONDENT BURDEN

Rachel A. Caspar and Paul P. Biemer, Research Triangle Institute
Rachel A. Caspar, RTI, P.O. Box 12194, Research Triangle Park NC 27709-2194

Keywords: NHSDA, interview length, experimental within-subjects design

I. INTRODUCTION

The application of principles from cognitive psychology to the survey research process has been shown to be beneficial, particularly in the area of questionnaire design. Questionnaire testing conducted in the cognitive laboratory often identifies problematic questions or instructions that were initially thought to be perfectly acceptable by the questionnaire designer. However, major drawbacks of the cognitive laboratory method have been the degree to which inferences can be drawn from a small group of subjects who may not be representative of the greater population of interest and who are interviewed in an environment unlike the one that will be encountered during primary data collection (see, for example, O'Muircheartaigh, 1999)

In an effort to address these drawbacks, RTI staff developed an approach for extrapolating results obtained in the cognitive laboratory to more closely represent what could actually be expected to occur in the field. This research was conducted during the development of the 1994 National Household Survey on Drug Abuse (NHSDA) questionnaire.

The NHSDA is the federal government's primary source of information on the magnitude of substance use and abuse in the United States household population. Conducted since 1971, the survey collects data by administering questionnaires to a representative sample of persons aged 12 and older at their place of residence. Since 1992, the survey has been administered by the Substance Abuse and Mental Health Services Administration (SAMHSA).

Due to the sensitivity of the data collected in the NHSDA, the study has always incorporated a methodology that includes self-administered data collection for the more sensitive topics. Interviewer-administered data collection is used for the more routine data that is collected. Up until 1999, the data were collected using paper-and-pencil forms. Respondents completed the self-administered sections by recording their answers on "Answer Sheets" that were never viewed by the interviewer. The answer sheet methodology was used to increase privacy and thereby encourage honest reporting of sensitive information.

In preparation for the 1994 NHSDA, a number of revisions were made to the content and structure of the questionnaire. An additional area of development was an attempt to provide illiterate, or semi-literate respondents with the same degree of privacy afforded to literate respondents. To this end, a new methodology for administering the NHSDA questionnaire to these individuals was developed and tested.

An early draft of the 1994 NHSDA instrument was tested in a small pretest and deemed too long by interviewers and respondents alike. The effect of this on the overall success of the NHSDA is multi-faceted. A questionnaire that is overly long will make it more difficult for interviewers to persuade sample persons to take part in the study. Once willing to participate, a respondent may terminate the interview because he/she grows weary of answering questions or provide less accurate responses toward the end of the interview due to growing fatigue or impatience. To determine the potential impact of the increased length of the 1994 NHSDA instrument it was necessary to collect more accurate, detailed information regarding the actual administration time for the revised NHSDA questionnaire.

In addition to concerns with the length of the interview, the pretest identified problems with the text developed for the interviewers to use with semi-literate or illiterate respondents. In the past, interviewers had been responsible for deciding whether respondents were capable of reading the questions themselves or whether the interviewer needed to read aloud for them. However, in situations where the respondent needed to have the questions read aloud, the interviewer was responsible for reading only the questions and not the answer categories. It was clear that more work was needed to develop a method for scripting answer categories that aided the respondents in answering the questions, while not unduly lengthening the time it takes to administer the survey. In addition, work was needed to develop alternative procedures for respondents who are able to read but simply prefer to have the questions read to them. These respondents may be willing to read the response categories for themselves.

II. THE RESEARCH DESIGN

The interviewer scripting was improved to make it more consistent and understandable throughout the instrument. Then, it was experimentally tested using two levels of detail to simulate two types of situations where it would be needed by the respondent (rather than reading it to all respondents regardless of reading ability).

Questionnaire content was given special consideration as the content affects the overall length of the instrument and the length of individual sections. Comparisons were made between the 1993 NHSDA questionnaire (which was being used at the time of this research) and the revised questionnaire to determine where additional questions had been added and where changes had been made to questionnaire items that had a significant impact on the data collected. Timing information from the experiment was collected on a section by section basis (although these data are not presented here); thus, we were able to determine which sections of the questionnaire were especially long and could work with SAMHSA to shorten these sections while maintaining the integrity of the data collected.

Our research objectives focused primarily on the issues of interview length, questionnaire administration, and questionnaire content. High priority was placed on the issue of interview length, and assuring that the final version of the revised questionnaire could be administered within the specified time constraints when either (a) the respondents read the questionnaire for themselves, (b) when the interviewer reads the entire questionnaire aloud to the respondent, or (c) when the interviewer reads the questions, but not the answer categories, to the respondent. Using laboratory interviews of a general sample of respondents, we sought to estimate the average length of time required to conduct an NHSDA interview under conditions (a) and (c) for the 1993 NHSDA questionnaire and under (a), (b), and (c) for the revised questionnaire. The ratio of the revised questionnaire laboratory completion time to the 1993 questionnaire laboratory completion time can then be applied to an estimate of the average completion time of a 1993 NHSDA interview in the field to arrive at an estimate of the interview length of the revised questionnaire in the field.

The use of laboratory interviews to estimate the average completion time ratio has two important advantages. First, due to the controlled laboratory setting, the variability among completion times may be reduced and thus, the experimental error can be dramatically reduced. This means that we can achieve acceptable precision in the timing estimates with only a small number of observations. Secondly, interviews can be monitored more easily by the project staff, and therefore, the causes of the completion time differences as well as other questionnaire problems can be more easily identified.

Another important feature of this experimental design was the use of repeated interviews with the same respondents. An advantage of interviewing each subject twice using a different questionnaire version or administration protocol is the reduction of experimental variation for the within-subjects comparisons. Also, by exposing the same respondents to two questionnaire versions or administration protocols, we obtain data, through debriefings, comparing one interview relative to the other. These data allowed us to determine respondent preferences for many critical design issues.

The two design factors included in the experiment are Questionnaire Version consisting of two levels: 1993 NHSDA (the instrument used in 1993) and 1994 NHSDA (the instrument under development for 1994); and Interview Administration consisting of three levels: respondent reads questions (R), interviewer reads questions and answer categories (FIQA), and interviewer reads questions only (FIQ).

For ease of explanation, we denote the 1993 version read by the respondent as the 1993:R and by the interviewer as the 1993:FIQ. Likewise, 1994:R, 1994:FIQA, and 1994:FIQ correspond to the revised instrument when: the questions are read by the respondent, both questions and answer categories are read by the interviewer, and only the questions are read by the interviewer, respectively. To minimize memory bias and other carryover effects, yet still allow for timely analyses, the second interview was scheduled to take place between 7 and 14 days from the first interview.

In order to test the effect of Interview Administration, it was necessary for us to interview nonreaders. However, due to the difficulty in identifying and recruiting a sufficient number of nonreaders to participate in the experiment, we developed a mechanism for encrypting the answer sheets in the FIQ and FIQA treatments such that even a literate person would need the aid of the interviewer to answer the questions. In this way, the effect of interview scripting on interview length and response could be more accurately determined. This process is described in greater detail below.

As can be seen in Table 1, our experiment involved 80 subjects (ten subjects per treatment cell). Each subject was to be interviewed twice for a total of 160 interviews. Based on our initial testing of the 1994 instrument, a sample of this size was calculated to be adequate to detect a difference of 15 minutes between completion times with a probability of Type I error of five percent and of Type II error of ten percent. This number of interviews was also adequate to investigate questionnaire content issues and other interview administration issues as appropriate.

Table 1. Treatment Assignment for the Within-Subjects Design

Condition	Questionnaire for 1st Interview	Questionnaire for 2nd Interview	Number of Subjects
1	1993R	1994R	10
2	1993FIQ	1994FIQA	10
3	1994R	1993R	10
4	1994FIQA	1993FIQ	10
5	1994R	1994FIQA	10
6	1994FIQA	1994R	10
7	1994FIA	1994FIQA	10
8	1994FIQA	1994FIA	10

III. IMPLEMENTING THE STUDY DESIGN

One of the flaws often cited with pretest work conducted in a cognitive laboratory setting is that subjects do not adequately represent the target population for the survey. This is generally due to the fact that subjects for laboratory work are recruited through some kind of "convenience" sampling. Signs may be posted in public buildings, or advertisements placed in local newspapers encouraging interested persons to call the researcher to schedule an appointment. The advantages of this method are that subjects are interested and usually can adapt their schedules to the needs of the researcher. It also can be a fairly inexpensive method for recruiting a large number of subjects. However, as noted earlier, the drawback can be that the recruited subjects differ considerably from the actual population of interest.

In an attempt to overcome this problem, we chose to recruit subjects for our experiment using a sampling procedure which would be more likely to target individuals similar to those interviewed as part of the regular NSHDA. To do this, we selected block groups in the Raleigh/Durham/Chapel Hill, NC area by a specially developed sampling scheme which balanced the sample on age and race characteristics according to NHSDA sample proportions for those characteristics. This list was sent to Survey Sampling, Inc. who provided us with the name, address, and telephone number for each housing unit in the block group (businesses were excluded).

The file delivered to us included 1,000 listings. From this file we randomly selected a subset of names to be among the individuals we attempted to recruit for Phase 1 of our experiment. A letter was mailed to each address describing the purpose of the study, the type of information we would be collecting and stated that a total of sixty dollars would be paid to each respondent who participated in both interviews. It closed by noting that a representative of RTI would call in the next few days to answer any questions they might have and to schedule a convenient time for the first interview to be conducted if the person was interested in participating. Approximately 41 percent of the sample scheduled an appointment to be interviewed. Refusals made up 32.3 percent and noncontacted cases were about 20.6 percent of the sample. We successfully recruited 78 subjects for Phase 1, with all subjects coming back for the second interview, for a total of 156 interviews.

Our Phase 1 experimental design called for using five different questionnaire versions. Development of the two 1993 versions was not difficult. Subjects receiving a 1993:R Questionnaire received exactly the same questionnaire that was used for the 1993 NHSDA. However, the interviewer instructions for conducting the interview were slightly different as will be discussed below. Subjects receiving a 1993:FIQ Questionnaire also were interviewed using the same document being used for the 1993 NHSDA. However, these subjects were provided with answer sheets with the questions in an encrypted form and the answer categories in English. These subjects were dependent on the interviewer to determine the content of each question, but could then read the answer categories for themselves. (Note: All encrypted items were created by translating the English text into Greek.)

Three questionnaire versions were created from the 1994 document. The 1994:R version mirrored the 1993:R questionnaire. All parts of the answer sheets were in English and the interviewer's instructions for conducting the interview were the same as in the 1993 instrument. The 1994:FIQ version mirrored the 1993:FIQ Questionnaire. All questions on the answer sheets were encrypted, while the answer categories were in English. Finally, the 1994:FIQA version is fully encrypted. Both the questions and the answer categories on every answer sheet were encrypted and thus the respondent was totally dependent on the interviewer to complete each section.

While similar to the "R" version, the "FIQ" versions employed one level of encryption. In this case it was essential that the respondent listen to the interviewer in order to understand each question on the answer sheets. Once the question had been read, the respondent could then read the answer categories silently and record the appropriate answer. As with the "R" version, the answer categories were not read to the respondent unless some particular question was raised.

The 1994:FIQA instrument utilized answer sheets which were fully encrypted. The interviewer was required to read both the question and the answer categories to the respondent. For this version it was necessary to provide the interviewer with a script with enough detail that the respondent would clearly understand how and where to mark each of their answers. A detailed script was also important to ensure that the interviewer would not have to look at the respondent's answer sheet to help them mark their answers, but would have all the information required within their own booklet.

The FIQ version may be of limited interest by itself, as it is unlikely that a field procedure forcing the interviewer to read every question on the answer sheets will be implemented. However, it was used in our analyses as an intermediate step in assessing the additional time added when a fully scripted instrument (the FIQA) in which the respondent is totally dependent on the interviewer was used.

Since the content of the 1993 and 1994 versions of the questionnaire were different, it was inappropriate to compare timing data for 1993 sections directly with 1994 sections. Therefore, our analysis focuses on the total completion time. In Table 2, contrasts between each pair of questionnaires tested in the laboratory are presented. These contrasts show the difference in total completion time (in minutes) for each pair of instruments.

We analyzed the timing data using analysis of variance. The model we assumed specified that completion time is a function of Questionnaire Version (V), Order (O: first or second), Interviewer (I), Subject or Person (P), and the Questionnaire Version by Order (V x O) interaction. There were significant differences in completion time among questionnaire version ($p<0.001$), and between first and second interviews ($p<0.001$). The Questionnaire Version by Order interaction was not statistically significant ($p<0.18$). Comparisons of time to completion among the five questionnaire versions, using the Student-Newman-Keuls test (with Experiment-wise $\alpha = 0.05$), indicated that the three different 1994 versions of the questionnaire were different from each other and different from the two 1993 versions, which did not differ from each other. Also, first interviews took an average of more than 15 minutes longer than second interviews (100.54 minutes and 85.46 minutes, respectively).

Given the differences in procedures between the present laboratory interviews and field interviews in people's homes, it is inappropriate to use the laboratory data directly to estimate how long the 1994 version of the NHSDA would take in a field setting. Instead, we developed a formula for estimating field times based on the experimental data as follows.

First, we obtained an estimate of interview completion time for field interviews completed during the first quarter of the 1993 NHSDA. This estimate was the best estimate available at the time the laboratory testing was conducted. The completion times from the preliminary 1993 NHSDA (denoted by F_{93}) indicate that interviews took an average of 65.72 minutes to complete, with a standard error of 0.658 minutes. We also found that 54 percent of respondents completed as many of the answer sheets by themselves as allowed by the interview protocol (we'll call these the Self-Completed interviews) and that 29 percent were entirely interviewer-administered (the Interviewer-Completed interviews). The remaining 17 percent of the sample completed only some of the answer sheets that were allowed for self-completion (the Mixed interviews). The times for the self-completed interview took about the same time (mean=64.40, s.e =1.26) as the interviewer-completed interview (mean= 65.71, s.e.= 0.71).

Table 2. Contrasts (In Minutes) Between Questionnaires Tested in Phase 1

	95%		95%
	Lower	Adjusted	Upper
Contrast	Limit	Diff.*	Limit
1993R - 1993FIQ	-13.94	-2.91	8.12
1993R - 1994FIQ	-36.53	-25.73	-14.93
1993R - 1994FIQA	-57.16	-48.22	-39.28
1993R - 1994R	-23.11	-16.75	-10.39
1993:FIQ -1994:FIQ	-31.67	-22.82	-13.79
1993:FIQ -1994:FIQA	-51.67	-45.31	-38.95
1993:FIQ - 1994:R	-22.77	-13.83	-4.89
1994:FIQ - 1994:FIQA	-28.67	-22.50	-16.33
1994:FIQ - 1994:R	0.18	8.98	17.78
1994:FIQA - 1994:R	25.18	31.5	37.72

*Differences and standard errors are computed on means adjusted for the unequal cell sizes.

This estimator of field interview completion time is not directly comparable to the laboratory estimates since the laboratory is a somewhat ideal setting for conducting an interview. What is needed is an estimator of the length of time required to conduct interviews using the 1994 instrument under field conditions.

To obtain such an estimate, we first constructed an estimator of the average completion time to conduct a set of laboratory interviews where **a***100 percent of the interviews are Self-Completed (as above), **b***100 percent of the interviews are Interviewer-Completed, and **(1-a-b)***100 percent of the interviews have Mixed administration. Let t_a, t_b, and t_c denote the laboratory completion times associated with Self-, Interviewer-, and Mixed- interviews, respectively, for one of the questionnaire versions (either 1993 or 1994). Then, the estimator of the average laboratory completion time for this set of interviews is $t = at + bt + (1\text{-}a\text{-}b)t$.

From the first quarter 1993 NHSDA data, **a**=0.54, and **b**=0.29. From laboratory interviews, we know t_a and t_b for both questionnaire versions, since these interviewing procedures correspond to the 1993:R and 1993:FIQ treatments, respectively, for the 1993 version and to the 1994:R and 1994:FIQA treatments, respectively, for the 1994 version The laboratory interview completion time for the t_c must be constructed by some combination of t_a and t_b, since this time is not directly observable from the laboratory study. Therefore, let $t_c = \psi * t_a + (1 - \psi) * t_b$ for Ψ between 0 and 1 and define the following ratio:

$$R_f = \frac{a*1994R + b*1994FIQA + (1-a-b)*[\psi*1994R + (1-\psi)*1994FIQA]}{a*1993R + b*1993FIQ + (1-a-b)*[\psi*1993R + (1-\psi)*1993FIQA]}$$

R_f represents the ratio of 1994 mean laboratory interview completion time to the mean 1993 laboratory interview completion time for a set of interviews with proportions **a**, **b**, and **(1-a-b)** of Self-Completed, Interviewer-Completed, and Mixed administration times, respectively, where the estimator of the Mixed administration time is $\Psi t_a + (1\text{-}\Psi)t_b$.

Finally, an estimator of the 1994 field interview completion time is:

$$\hat{F}_{94} = R_f * \hat{F}_{93}$$

where $\hat{F}_{93}$ is the 1993 field estimate of 65.72 minutes. To estimate $\hat{F}_{94}$, we used values of Ψ=0, .5, and 1.0 to provide a lower bound, midpoint, and upper bound, respectively, for t_cb for 1993 and 1994.

The estimate for $\hat{F}_{94}$ for Ψ = 0.0, 0.5, and 1.0 were 92.65, 90.50, and 88.34 minutes, respectively. Based these findings, we estimate that using the 1994:R and 1994:FIQA procedures in the field would produce an average interview completion of approximately 91 minutes if the proportions of Self-Completed, Interviewer-Completed, and Mixed interviews found in the preliminary first-quarter of 1993 hold.

Based on the results obtained in the first phase of testing, revisions were made to the 1994 NHSDA Questionnaire. For the most part, revisions were made to shorten the instrument as the length from the first round of testing was deemed to be too long. Our goal was to keep the 1994 NHSDA to as close to one hour as possible. Following these revisions, a second phase of testing was initiated. We used the same procedures as in Phase 1 with a few modifications. We tested two versions of the 1993 questionnaire that were identical to those used in Phase 1 (1993:R and 1993:FIQ). Two versions of the revised 1994 instrument were also tested (1994:R and 1994:FIQA).

A total of 41 subjects were included in this phase of testing. These subjects were recruited using the same procedures described earlier. Each subject interviewed was randomly assigned to one of the following four treatment cells: 1993:R - 1994:R; 1994:R - 1993:R; 1993:FIQ - 1994:FIQA, and 1994:FIQA - 1993:FIQ. All interviewing procedures were the same as those used in Phase 1.

In Table 3, contrasts between each pair of questionnaires tested in Phase 2 are presented. These contrasts show the difference in total completion time (in minutes) for each pair of instruments. We again analyzed the timing data using analysis of variance. The model we specified was the same as that used in Phase 1.

Table 3. Contrasts (In Minutes) Between Questionnaires Tested in Phase 2

Contrast	95% Lower Limit	Adjusted Diff.*	95% Upper Limit
1993:R - 1993:FIQ	-5.89	1.92	9.23
1993:R - 1994:FIQA	-368.95	-30.15	-21.35
1993:R - 1994:R	-7.98	-1.97	4.04
1993:FIQ - 1994:FIQA	-35.67	-29.37	-23.07
1993FIQ -1994:R	-5.13	2.09	9.31
1994:FIQA -1994:R	21.83	30.83	39.78

*Differences and standard errors are computed on means adjusted for the unequal cell sizes.

It should be noted that the within subjects design for Phase 2 was an incomplete factorial design since the 1993:R-1994:FIQA and the 1993:FIQ-1994:R combinations were not administered. This was done because our interest was in comparisons of parallel versions of the questionnaire. This design made such comparisons a within-subjects factor. Any comparison of other versions of the questionnaire are between-subjects comparisons. As in Phase 1, the order of administration of questionnaire versions was randomized and balanced within subjects.

There were significant differences in completion time among questionnaire versions ($p < 0.001$), and between first and second interviews ($p < 0.001$). The Questionnaire Version x Order interaction was also statistically significant ($p = 0.03$).

Comparisons of time to completion among the questionnaire versions, using the Student-Newman-Keuls test (with Experiment-wise $\alpha = 0.05$), indicate that the 1994:FIQA version of the questionnaire took significantly longer to complete (102.04 minutes) than the other three versions (73.02 minutes for 1993:FIQ, 67.55 for 1993:R, and 69.54 minutes for 1994:R), which did not differ from each other. Also, first interviews took an average of 9.8 minutes longer than second interviews (82.57 minutes and 72.80 minutes, respectively).

Using the same procedures described above, we calculated an estimate of the length of the 1994 NHSDA instrument in the field using the results of this second laboratory experiment.

New estimates of field interview length were computed as 79.25, 77.17, and 75.06 minutes for Ψ = 0.0, 0.5, and 1.0, respectively. Thus, based on our laboratory findings, we estimated that using the 1994:R and 1994:FIQA procedures in the field would produce an average interview completion time of approximately 77 minutes if the proportions of Self-Completed, Interviewer-Completed, and Mixed interviews found in the first quarter of 1993 held for 1994 as well. While these estimated times were still over our goal of a one hour interview, they were much shorter than the estimates from Phase 1.

IV. ASSESSING OUR ESTIMATION STRATEGY

To determine whether our laboratory testing and our subsequent estimation strategy were accurate, we recently revisited the 1993 and 1994 NHSDA data to recalculate our estimates based on complete field results from 1993. The 1993 data we used to calculate our original estimates came only from the data collected during Quarter 1 of 1993. We can now use the results from the full year of data collection. The average interview length did not vary dramatically; the average interview length for 1993 was 63.43 minutes (the Quarter 1 average was 65.72 minutes). However, the proportion of cases that were completed in each of the three interview modes (completely by the respondent, completely by the interviewer, or some mixture) did change significantly. For the year, the proportions were .775, .132, and .092 respectively. Using the estimation strategy described above, we generate the estimates for the 1994 NHSDA interview length were 71.04, 69.77, and 68.50 minutes for Ψ = 0.0, 0.5, and 1.0, respectively.

In fact, the 1994 NHSDA averaged 68.37 minutes. This average is quite close to our estimate where Ψ = 1.0. The reason for this can be found in the fact that the proportion of 1994 cases completed in each mode varied significantly from our actual 1993 experience. In 1994, 96.7 percent of cases were administered completely by the respondent, 1.2 percent were administered completed by the interviewer, and 2.1 percent were administered in some mixture. These percentages would be most closely reflected by our upper estimate which assumes the largest number of cases are completed under the shortest mode of administration.

From these data, it appears our estimation strategy was fairly accurate. The large difference in the administration percentages was unexpected at the time, but in retrospect is not surprising. Since much of the redesign work for the 1994 NHSDA instrument involved developing a methodology to allow a larger number of respondents to complete the interview on their own, it is not surprising that the percentage of cases completed solely by the respondent would rise.

ACKNOWLEDGMENTS

We would like to acknowledge the following individuals who were critical to the success of this study: Teresa Edwards, Michael Hubbard, James Kennedy, and Kelly Wayne.

REFERENCES

O'Muircheartaigh, C. (1999). "CASM: successes, failures, and potential," In Sirken, et al. (eds.), *Cognition and Survey Research*, Wiley, pp 39 - 62.