Survey Methodology, December 2000
Vol. 26, No. 2, pp. 199-207
Statistics Canada, Catalogue No. 12-001

Screen Design and Question Order in a CAI Instrument Results from a Usability Field Experiment

Marek Fuchs [1]

Abstract

Screen design and questionnaire design affect the interviewer behavior in a CAI environment. Previous research has shown that interviewers can work more properly and efficiently if suitable functions and features are incorporated in the CAI instrument. Usability experiments with the household roster of two large government surveys have shown that using grids and tables is an important feature to facilitate the interviewer's performance. While these experiments were conducted under laboratory conditions, we have results from a first field experiment. In March of 1998 a CATI survey on immigrants was fielded in Germany (response rate 84%, n = 501). Four different versions of a household roster were compared in this production study, testing two different screen designs together with two different question orders in a 2×2 factor design. The four versions were randomly assigned to interviewers and respondents. Time measures were built into the CATI program, and 234 randomly selected interviews were video taped and analyzed according to a coding scheme. Based on the data we assessed the usability of different CAI design features. The results show that the screen design as well as the question order have a significant influence on interview duration and interviewer behaviors. Especially the grid based and topic based version allows the fastest performance in terms of time used to complete the instrument. Results from the coding data suggest that the differences between versions are due to specific interviewer and respondent behaviors. The data indicates that the grid based topic version enables a respondent oriented interviewer behavior, and thus allows the best interviewer performance in terms of duration.

Key Words: Computer assisted interviewing; Usability Testing; Field experiment; Screen design; Question order.

1. Introduction

Computer assisted interviewing is on its way to becoming a standard survey technique (Couper, Baker, Bethlehem, Clark, Martin, Nicholls and O'Reilly 1998). In telephone surveys as well as with personal interviews, more and more studies are conducted using computer assisted interviewing techniques (CAI). Many of the large government surveys in the US are in the transition to CAI or have completed it already. Even in Europe, we observe a shift towards computer assisted interviewing (Schneid 1991; Fuchs 1994, 1995; Laurie and Moon 1997; Projektgruppe SOEP 1998) - even though, the methodological aspects of this development do not constitute the main focus of European research, so far.

Researchers and people responsible for fielding surveys rely on computer assisted interviewing for several reasons: (Sometimes it seems, however, that substantial arguments are less important than just a specific market rush towards CAI.)

- They hope to collect data of higher quality due to built-in consistency checks and range checks during the course of the interview.
- CAI provides the possibility to use automated skip patterns and allows to design more complex instruments without putting too much burden onto the interviewers.
- They hope to spend less time and money for interviewing and post-processing and decrease survey budgets once the up-front investment for hardware and software is payed off.
- They hope to benefit from CAI's ability to read external data into the interview which is especially interesting with panel studies.

The general movement towards CAI is evaluated positively. Researchers and field directors benefit from it (Nicholls and deLeeuw 1996) and interviewers (Couper and Burt 1994) as well as respondents (Baker 1992), reveal a great deal of sympathy or at least acceptance. On the other hand, computer assisted interviewing has introduced some additional problems into the interview situation, too: in the early years methodological research was mainly concerned with hardware and software problems (see Couper, Groves and Kosary 1989; Weeks, 1992 for overviews). Instead, recent studies dealt with interview and respondent acceptance, interview duration, and usability issues (Couper *et al.* 1998 for an overview). The present paper contributes to this later discussion of "technology effects" (Fuchs, Couper and Hansen 2000).

2. Theoretical Background

For the purpose of the following analysis the theoretical focus is mainly on two usability issues: (1) segmentation of the interview flow and (2) lack of interviewer flexibility.

1. Dr. Marek Fuchs, Catholic University of Eichstätt, Department of Sociology, Ostenstrasse 26, 85071 Eichstätt, Germany. E-mail: marek.fuchs@ku-eichstaett.de.

1. Segmentation: in a CAPI environment the interviewer has an additional burden: the process of keying takes place in the interview situation. Usually, an interviewer reads a question, receives an answer, enters the data, presses [enter] and then the next screen with the following question appears. Compared to PAPI interviewers cannot look ahead and anticipate the next upcoming question while recording the answers to the previous one and they cannot start reading the next question before pressing [enter] - they cannot work simultaneously on both tasks. As a result of this procedure the interviewer respondent interaction is segmented by [enter] keys. So far we do not have quantitative evidence that this kind of segmentation harms the data or the interview situation. But it is argued that the interviewer loses the "big picture", and the relevance of questions and their relationship to each other may be unclear (House 1985; Groves and Mathiowetz 1984).

Our findings from several series of usability tests in the lab concerning the screen layout of a household roster (Couper *et al.* 1997; Hansen, Couper and Fuchs 1998) led to the suggestion of a specific screen layout that allows the interviewer to develop a more complex understanding of the instrument, maintain the interaction with the respondent, and enter data at the same time: Two different versions of a series of questions were tested under laboratory conditions in terms of the time necessary in order to complete the questions and ease of use. We compared a so-called item based design with a grid based design. House and Nicholls (1988) distinguished between three approaches in screen design for computer assisted instruments: item based, screen based and form based design. In the item based approach one question and one input field are displayed at a time, and logic operations are performed in the transition from one item based screen to the next. This design is easy to program and focuses the interviewer's attention on the actual question. The screen based approach combines several items that need to be answered in sequence. All logic operations are executed after each item. On a form based screen, many items are presented at the same time in a table or grid and the interviewer may use the cursor keys to move from field to field and to complete them in any order.

The item version tested in our experiment matches the characteristics specified by House and Nicholls (1988) for a screen based approach. In contrast, the grid based design is best described as a form based instrument. It allows interviewers to record the information in the order chosen by the respondent, it provides the interviewer with a better overview of the instrument and it more easily allows updates and backups (for details see Couper *et al.* 1997). Also, the design matches the interviewers' demand for more questions on one screen - both for speed of administration and for context knowledge. The following graph gives an impression of an item based and a grid based CAI screen design.

We found evidence that the grid based design reduces the segmentation: interviewers could start reading the next upcoming question while still entering the data to the previous question. Even backing up seems to be easier within a grid design. On the other hand, we found only modest support for a grid based design in terms of time used to complete the task (for details see Couper *et al.* 1997). This leads to the question: what can we do to decrease segmentation and to further improve the efficiency of a household roster in terms of duration?

2. Lack of flexibility: The second feature that might cause problems in a computer assisted interview is the lack of flexibility. One of the advantages of a CAI instrument is the fact that an interviewer can hardly skip any questions. Although CAI instruments can make extensive use of skip patterns and filters, they apply a predefined question order. Usually, each question needs an [enter] key before the system goes on to the next screen. It is seen as an advantage that this rigid question order avoids any trouble the interviewer might have with the routing through the instrument, questions for specific respondents, filters and skip patterns and so on. He or she can abandon this task and focus on the administration of the actual items. On the other hand, this causes a very strict question order and provides the interviewer with little flexibility in terms of question order. A small example demonstrates this effect: most CAI instruments apply a question order to their household roster, where all items for one person are asked before the interviewer works through the same items for the next person ("person based design" see Couper *et al.* 1997; Fuchs 2001 or "grouped questions" see Moyer 1996). The CAI instrument, for example, might request the respondent's age, educational level, and other questions first before asking for the age of the respondent's wife. (This can be explained in part by the way computer programs and data bases work: households represent the main records and persons or other entities are treated as subrecords.) When completing the questions of a household roster it might happen (and in fact it happens quite often, see below) that the respondent provides not only the answer to the current question (*e.g.*, "I'm 34 years old") but also to a related question: "I'm 34 years old and my wife is 32 years old" or the respondent might answer "We are all Black" when asked about his or her own race (Oksenberg, Beebe, Blixt and Cannell 1992).

While working with a paper instrument it is an easy task for an interviewer to make immediate use of the additional information provided by the respondent. In case he or she answers, for instance, "We are all black" the interviewer can easily mark the appropriate check boxes for all household members at once. For someone interested in questionnaire design this leads to the following question: given the lack of flexibility in a computer assisted environment, what is the best question order for collecting information about all household members?

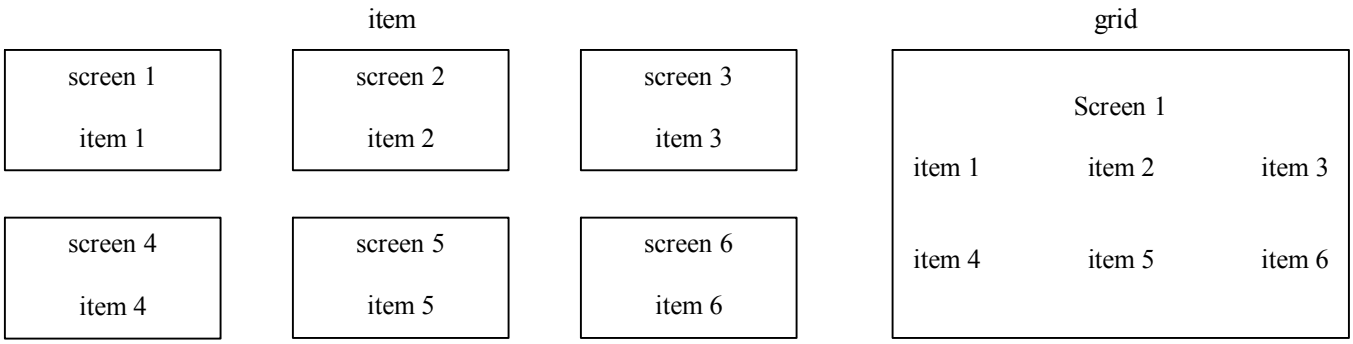

Figure 1. Item Based Design *vs*. Grid Based Design.

Moore and Moyer report results from an experiment on two different question orders designed for collecting information about all eligible persons in a household (Moore and Moyer 1998a, 1998b). The first question order asks all questions for the first eligible person in the household and moves on to the next person, when all questions are completed. This question order is called a person based approach. In the second version, the topic based approach, the first question is asked for all eligible persons, then the second question for all persons and so on. Moore's and Moyer's results show strong support for a topic based design: the topic version leads to less item non-response, less break offs and refusals and is substantially shorter. Besides interviewers show significant preference for this version.

In the experiment presented in this paper we tried to make use of the advantages of a topic based approach and of a grid based screen design: we combined the two screen designs (item based design *vs.* grid based design) with the two question orders (person based order *vs.* topic based order) and tested all four resulting versions in a field experiment. In doing this, we had the following assumption in mind: the usability of a CAI instrument is not only a programming issue, but it is also connected to the questionnaire design and to the interview as a social situation. Both aspects of a computer assisted instrument, its screen design and its question order, support or hinder a smoothness of the interview flow. Based on the results of the previous research we had the following hypothesis: The combination of a grid based screen design and a topic based question order allows the most efficient interviewer respondent interaction.

3. Methods

The experiment took place in Germany in March 1998. Immigrants of German origin from Poland, Rumania and the former Soviet Union were surveyed. Starting February 28, 1998 and ending March 20, 1998 15 interviewers completed $n = 501$ interviews. All respondents received an advanced letter and were called by phone up to 15 times. The response rate reached 84% and item non-response was considerably low. The interviews were conducted using the CATI program CI3. About 95 questions on various topics were asked. The average interview lasted 23 minutes.

Four versions of a small household roster with three items per person were included in the instrument: an item/person version, a grid/person version, an item/topic version and a grid/topic version. All versions applied the same question wording and interviewer instructions, however, we modified the screen design and the question order according to the theoretical approach mentioned before (Figure 2). The item based person version is considered to be the standard version - it represents the questionnaire design usually applied to socio-demographic portions in CAI surveys. One of the four versions was randomly assigned to each interview - and thus to interviewers and respondents. We measured the total time needed for the household roster and in addition the time spent on each single item in that section of all 501 interviews. In addition, 234 interviews were selected at random and the interviewer working through the household roster section was video-taped. The video segments were coded in terms of interviewer behavior and respondent behavior and the resulting data was combined with the time measurements.

4. Results

The durations of the four versions differ significantly from each other: interviewers needed 6.6 seconds per item in the item based person version (which is considered to be the standard one). In contrast each item took 5.5 seconds in the grid based topic version. This is a reduction of about 17% for the grid based topic version. The two other versions are in between.

It is important to mention that both factors seem to contribute to the decrease in time used to complete the task. If we distinguish between the two factors, we end up with the following results: the two topic based versions are significantly shorter than the two person based versions and the two grid based versions take significantly less time than the two person based versions. The combined effect applies to

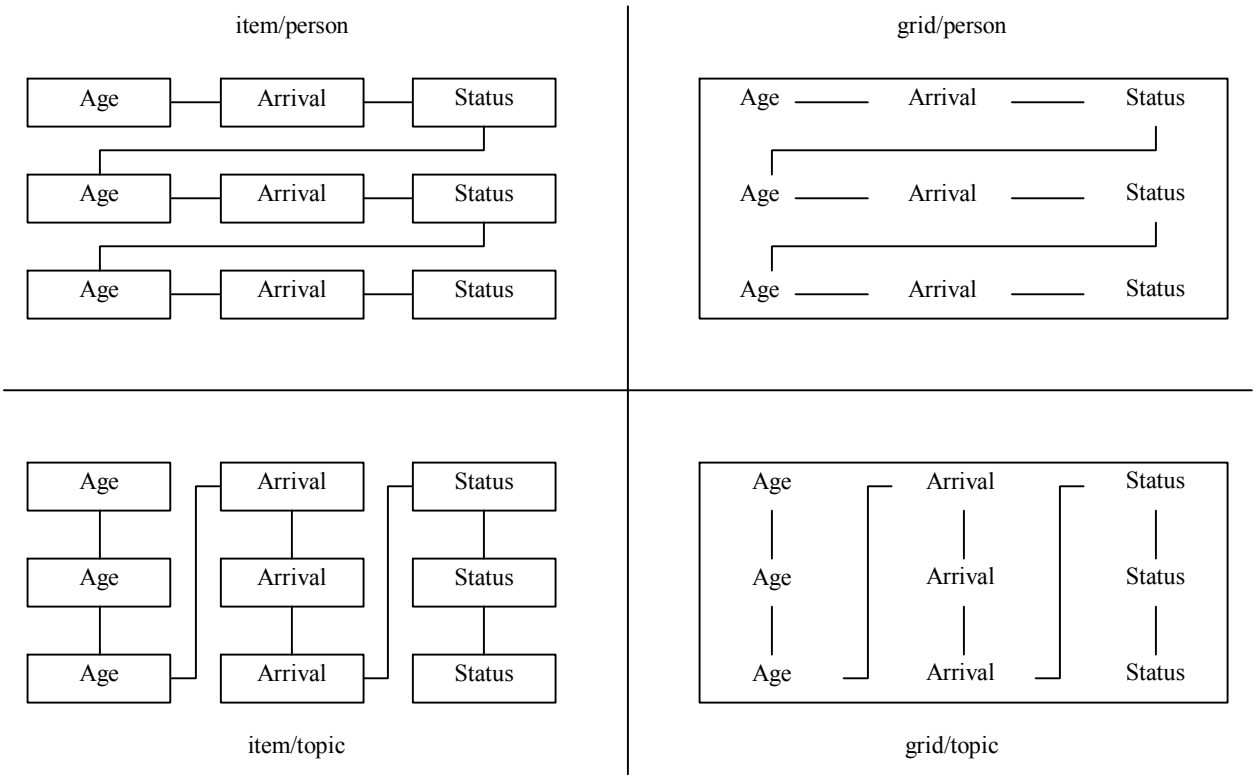

Figure 2. Four Versions Tested in the Experiment (Each Box Represents One Screen).

the grid based topic version and leads to the value of 5.5 seconds per item. (An analysis of variance reveals that both factors - the screen design as well as the question order - contribute independently to the decrease in time (screen design: $p < 0.01$, one third of total effect; question order: $p < 0.001$, two thirds of total effect, no significant interaction).)

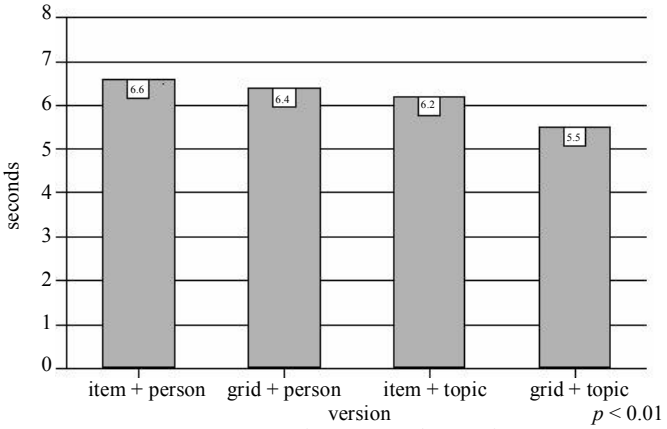

Figure 3. Duration per Item by Version.

But why is the grid based topic version faster? A detailed analysis shows that this version is especially faster when collecting the information for the second and all following persons in the household - a significant impact, that is called loop effect (Fuchs 2001). This term describes the following phenomenon: the interviewer takes much longer to collect the information for the first person in a household compared to all subsequent persons. The average loop effect sums up to 3.4 seconds per item which is a reduction of about 38% compared to the first person (Table 1).

The loop effect is not specifically characteristic for this experiment. We recognized loop effects in our previous experiments with the NHIS household roster, too (Couper *et al.* 1997). It is, however, interesting to observe that the loop effect is significantly larger for the topic based versions than it is for the versions that follow a person based question order (Table 1). Thus the topic based versions do increase the acceleration for the second and all subsequent persons in a household and consequently show a larger loop effect. (One implication of our experimental design might be that interviewers did not know what version they were approaching. This may have decreased their performance on the very first item. But this effect should be the same across all versions, so the results should not be affected.)

Table 1
Duration and Loop Effect (Seconds)

Item	Age	Arrival	Status	All items
Duration per item				
First person in household	9.4	9.9	7.7	9.0
All other persons in household	6.6	5.7	4.5	5.6
All persons	8.0***	7.8***	6.1***	7.3***
Loop effect				
Differenz between first and all other persons in the household	-2.8	-4.2	-3.2	-3.4
Loop effect by version				
Grid + topic	-6.8	-6.4	-4.6	-5.9
Item + topic	-5.2	-8.3	-4.3	-6.0
Grid + person	-0.5	-2.7	-3.0	-2.1
Item + person	0.3	-0.3	-1.3	-0.4
Average loop effect	-2.8***	-4.2**	-3.2**	-3.4***

**$p < 0.01$; ***$p < 0.001$

Analyzing the video tapes we can provide reasons for these differences at least in part: given the topic based conditions, both interviewers and respondents adapt differently to the interview situation compared to the person based versions. When asking the questions for all persons in the household, the respondent recognizes the logic of the procedure very quickly. In quite a high proportion of all cases (about 30%) their reaction to this is "We all arrived in the same year" (meaning: "Don't ask me this question again and again").

If the instrument follows a person based design, the interviewer has to memorize this piece of information, and if it comes to the next person, he or she needs to remember: "Do not ask this question again, the respondent gave you the appropriate answer already!" Only in a few cases they really do, most of the time they just ask the question again. This is especially true when using an item based screen layout that gives no clues in terms of the answers to the same question for the other household members. In a topic based design instead, the interviewer can easily adapt to that situation. Thus he or she just enters the same code for all persons in the household without asking the question repeatedly. Both the interviewer and the respondent get used to the questions, and so the question answer process runs with less verbal contributions from the interviewer's side as well as from the respondent's. Both interviewer and respondent can anticipate the next question and the interview runs more smoothly. This is especially true when the CAI instrument makes use of a grid and provides further context information, *e.g.*, the responses for other household members to the same question. (Looking at the results reported in the lower part of Table 1 we conclude that the grid based person version does not benefit to the same extent from the advantages of the topic based approach. However, due to the grid design the loop-effect is considerably larger than in the item based person version.) As a result the time used per item is substantially shorter and the interviewer can provide respondent oriented interviewer behavior similar to Schober and Conrad's (1997) findings.

Providing feedback by the interviewer sometimes works as a signal that he or she has recorded the answer to the previous question in order to stimulate the respondent, so that the latter guesses about the next question and reveals the appropriate answer even without an additional stimulus. In extreme this might lead to a respondent behavior where he or she provides the information about all persons in the household at once: "We all came in the same year". The different versions tested in this experiment impel and support such behaviors to different degrees. From our results we can conclude that the grid based topic version stimulates interviewers and respondents to deviate from the scripted interview to a higher degree than the other versions. As far as duration is concerned this version allows the interviewer to make efficient use of information provided for all household members at once. Evidence from the video coding support our interpretation of version-specific occurrences of time saving interviewer behaviors (1) and respondent behaviors (2):

1. By means of analyzing the video tapes we observe quite a lot of interviewer behaviors that do not follow standard interviewer procedures: besides the fact that about 78% of all items are read as worded, interviewers do not administer 9.3% of all items to the respondent. In another 5% of instances, the interviewer does not read the question but instead provides a different stimulus containing the relationship of the next person to the respondent (*e.g.*, "... and your wife?"). (It is interesting to recognize that interviewers chose the same verbal expressions on their own that Moore and Moyer 1998a, 1998b scripted in their experiments on question order.) In 5.5% of all cases the interviewer does not read the question but rather verifies the answer ("... and your wife is 32 years old?"). Some incomplete questions and wrong fills are observed, too. In total we have about 22% of all items affected by at least one interviewer behavior that does not follow a standardized interview script - which is a surprisingly high value considering that all interviewers were aware of the fact the interviews were video taped! Compared to other studies on interviewer behavior, however, the values are considerable lower. For example Oksenberg, Cannell and Blixt (1996) applied behavior coding to the National Medical Expenditure Survey and reported 37% to 41% of such interviewer behaviors. We will come back to the question of whether or not these behaviors help obtain valid measurements.

We draw the following conclusion from these particular findings: most of these behaviors indicate kind of a short-cut, *e.g.*, the interviewer does not read the question text as worded, he or she tries to make the conversation smoother and more suitable in terms of conversational rules. From our point of view this indicates that interviewers do not want to ask for information the respondent provided already. They do not want to behave unresponsively toward the verbal contributions of the respondent, instead, they wish to follow conversational rules. As a side effect these behaviors are less time consuming than standard interviewer behaviors. In our perspective, the priority therefore lies not with saving time, but with customizing the question answer process to respondent behaviors not anticipated and not absorbable by the computer assisted instrument.

In order to compare the four screen design versions in terms of the degree of interviewer deviations from the standard interview script we have computed the proportion of items per case affected by this kind of behavior. Large differences in interviewers not following the scripted interview between the four versions are to be noticed: Applying the grid based topic version to an interview results in more than twice as many such behaviors (the average proportion

of items affected is 0.48) than the item based person version (0.21) which is the standard for most studies so far. (An analysis of variance indicates that both factors contribute independently to the overall effect (screen design: $p < 0.001$; question order: $p < 0.001$; no significant interaction effect). About 25% of the overall effect can be attributed to the screen design, about three quarter to question order.) And this contributes to the time used for interviewing: items affected by a interviewer behavior not scripted in the interview take substantially less time (4.0 seconds) than the regularly administered items (6.8 seconds; $p < 0.001$).

Table 2
Interviewer Behavior and Respondent Behavior by Version

	Grid + topic	Item + topic	Grid + person	Item + person	Total
Average proportion of items affected by interviewer behavior not following the scripted interview per case	0.48	0.43	0.34	0.21	0.36***
Respondent provides information for all persons in the household at once	38.2%	44.4%	29.0%	10.8%	29.7%***

***$p < 0.001$
In order to differentiate between the proportion of cases affected by a certain respondent behavior and the average proportion of items per case (!) affected by a certain interviewer behavior we used percent notation for the first and decimal notation for the later.

2. Additionally an analysis of the respondents' behavior shows that the topic design leads to a higher proportion of cases (42,3% compared to 19.7% for the person approach; $p < 0,001$) where the respondent provides at least once in the household roster section the information for all persons or a group of persons at once (*e.g.*, "We all came in the same year"; "We all have the same legal status"). By contrast, the difference of the grid based design from the item based design is considerably smaller (33,6% *vs.* 26.1%) but does not reach the level of significance. However, an analysis of the interaction reveals a significant interaction effect ($p < 0.05$): Using a topic oriented question order the grid design does not make a significant difference. However, on top of an topic oriented question order the grid design increases the number of instances where the respondent provides the information for all household members at once.

It is surprising that results differ even for the two screen designs when using a person oriented question order. The study was administered by telephone, the respondents not being aware of the screen design at all. The only possible explanation is based on the fact that the interviewers modify their behavior in concordance with the screen design, stimulating the respondent differently. Accordingly, respondents, as well as interviewers, react to the screen design and the question order under the grid based person design in a way that facilitates the interviewer respondent interaction and thus helps smoothen the interview flow. (As seen before, the interviewers change their behavior even under the grid based person condition (Table 2), however, the question order does not stimulate respondents to behave accordingly.)

One possible drawback of these interviewer and respondent behaviors might be a lack of data quality due to changes occurring in the predefined question answer process; instead, the respondent considers the answer less intensively and thoroughly. We observe only very few item missing values so an analysis of this standard indicator for data quality is not efficient. In fact, we do not expect a higher proportion of item missing values in either version. One might, however, be concerned about the homogeneity of the answers provided by the respondent. In a high proportion of cases he or she listens to the full question text only once and that could contribute to a less thorough consideration when answering the same question for subsequent household members. Additionally, answering for all household members at once ("We all arrived in the same year") might increase the homogeneity of the response and thus decrease data quality.

Table 3
Average Number of Different Categories (Homogeneity) per Household by Version

Variable	Grid + topic	Item + topic	Grid + person	Item + person	Total
Year of arrival (19 categories)	1.2	1.2	1.2	1.3	1.2
Status (4 categories)	1.3	1.3	1.3	1.3	1.3

No significant differences

In order to assess this possible drawback we computed the number of different response categories chosen by the respondent on a particular item for all household members (*e.g.*, for year of arrival: respondent 1985, partner 1987, daughter 1987, son 1988 = 3 different response categories). This should give us an idea of whether or not only those respondent make use of the short-cut ("We came all in the same year") for whom this is actually valid, or whether even other respondents provided one answer for all household members even though they should have chosen two or more different response categories because of the situation in their particular household (unfortunately we have no external validation for the responses provided). In looking at the average number of different response categories (Table 3) we do not notice any differences in terms of homogeneity of data. For the year of arrival as well as for the legal status (as a German or a foreigner) there is no visible difference between the versions. For all versions the average number of different response categories chosen (one for each person in a household) shows no significant difference.

These finding provide only weak evidence that a grid design does not harm data quality. Other standard data quality indicators need to be assessed with larger data sets

in order to decide whether or not data quality is affected. However, based on the data available, we are unable to proof a effect on the validity of the responses.

5. Discussion and Conclusion

Our results from a comparison of four versions for a household roster (using the same question wording across versions) indicate that interviewers as well as respondents perform more efficiently under the grid based topic condition than with the other three versions. Combining a grid based screen design and a topic based question order reduces the average duration by about 17%. Two thirds of this reduction can be attributed to the question order, approximately one third to the screen layout. It is important to mention that the effect of the screen design is less pronounced than the one of question order and - compared to the effect on duration - even smaller on interviewer behavior and respondent behavior.

Even though the effects of the grid design on interviewer behavior and respondent behavior are far from large, they help to elicited two reasons for the better performance of the grid based topic version in terms of interview duration: (1) in the grid based topic version, the interviewer as well as the respondent adapt better to the logic of the question answer process, both anticipate the next question more easily and the question answer process runs more smoothly. (2) This version leads to more occurrences in which the respondent provides the information for the persons in the household faster and more often the respondents reveal the information for all household members or at least for one group at once. Even though the results are not fully consistent, this particular version makes it easier for the interviewer to adapt to this situation, record the information and stimulate the respondent to give the next appropriate answer without repeating the full question text.

Our findings contribute to the discussion of how to design survey instruments for interviewer administered computer assisted data collection. Based on the results reported in this paper we can draw the conclusion that making use of grids facilitates the interviewer respondent interaction and helps speed up data collection. Our experiments on item design *vs.* grid design conducted in the University of Michigan Survey Research Center's usability laboratory have shown that we can improve interviewer performance by providing grids (Couper *et al.* 1997). Moore and Moyer (1998a; 1998b) have demonstrated that one can improve interview efficiency by switching to a topic based question order, too. The present paper indicates that the interview situation benefit even more when combining both features.

Using grids and a topic based question order causes a greater amount of instances where the interviewer deviates from the scripted interview. From a rigid methodological point of view this might be seen as an important drawback, especially, if the interviewer deviates from the standard interview script using global questions for all household members. For example, Martin (1999) showed a significant increase in the number of people enumerated in a household if extra questions were asked. In addition, Kindermann and colleagues (1997) demonstrated for non-household roster type questions, that additional cues on victimization significantly increase enumerations of crimes. Generalizing these results to global questions across persons, one would expect a decrease of data quality as interviewers use non-scripted behaviors that apply global questioning methods. However, the results reported in this article do not indicate that interviewers are using global questions and the author does not recommend to make extensive use of global questions when designing a survey instrument. Instead, the findings lead to a screen design that allows interviewers to make use of information reported when the respondent switches to a global mode and provides the information for several household members at a time. So, we do not want to encourage researcher to make extensive use of global questions and we do not want to see interviewer modify the scripted questions in order to ask global ones. However, when confronted with a respondent providing more information than actually asked for, the screen design of the CAI instrument should not prevent interviewers from making use of it.

A grid based design has been proven to facilitate the interviewers job with respect to this task, because it allows interviewers to adjust their behaviors in concordance with general conversational rules. Basic findings of behavior coding suggest that interviewers frequently deviate from specific interviewing procedures. "These changes often reflect adjustments made by the interviewers to meet the exigencies of the situation: to melt it more congenially with communications immediately preceding it, or to adjust to the respondent's particular situation" (Oksenberg *et al.* 1992: 3). This is especially necessary when respondents do not limit their answers to the information requested by the question, but elaborate it or provide additional information. "Avoiding the appearance of not paying attention to the respondent, interviewers in this situation frequently filled in the answer themselves without asking the question, or asking it only in part" (Oksenberg *et al.* 1992: 5). They thus try to switch to more respondent oriented procedures to avoid looking unresponsive. A grid based screen design and a topic oriented question order supports interviewers to interact according to these conversational rules and with respect to the interview situation's needs. This might be acceptable or even preferable as long as we are talking about factoid questions and as long as these interviewer behaviors do not harm data quality (*e.g.*, leading question or probes).

What needs to be done in order to improve the computer assisted instrument in its supporting function for the interviewer respondent interaction: Our data suggests that the grid based topic version leads to a specific interview flow,

so that interviewer and respondent can easily adapt to it. Jeff Moore (1996; Moore and Moyer 1998a, 1998b) has shown that interviewers prefer the topic based version. By contrast, we know little about the respondents' satisfaction with that question order. Assessing their opinion about the different version is consequently an important goal. Moreover, we do not know whether this version matches the way in which information is stored in the respondents' brains. It might be, that respondents can easily adapt to this version, but that in terms of cognitive and social burden or in terms of correctness of answers it is not the right method. Additionally, we need to focus on the question whether or not we can transfer our findings from a household roster to other segments of a questionnaire. Right now we are conducting a series of field tests comparing different design solutions for factoid information other than household roster information and for attitude items. The versions tested in this experiment differ in the degree of contextual information provided to the interviewer while administering a particular item (previous questions, next questions *etc*.). The general question sounds: what happens if we use grids or form based screens more extensively? Under what conditions and circumstances does it help to improve interview efficiency and what are the limitations to this approach? However, it is too early to present any results at this time.

In addition, there are more unanswered questions that need to be addressed in future research. Personally I would like to suggest a specific approach to assess these questions assuming that computer assisted instrument design is of importance to different clients: researchers, interviewers and respondents. Of course, it is important that a CAI instrument meets the researcher's needs to obtain his or her measurements and also that the question answer process be well designed for each single item. However, in my view considering the social dimension of the interviewer respondent interaction and the behaviors in between single items is also a matter of importance. If the CAI instrument disturbs the social dimension of the measurement process it might harm even data quality. So far we do not know which approach allows the best compromise between validity and reliability of the measurement process on the one hand and a smooth short and non-embarrassing interview flow on the other hand. In order to find out to what respect a specific CAI screen design might harm data quality and how it helps save time, money and interviewer effort we need to conduct more usability studies.

To assess the questions mentioned above we do need more field experiments. Due to the fact that we want to analyze the social dimension of the interview and its effects on interviewer behavior as well as on interview duration, laboratory experiments do not meet our needs completely. Of course laboratory experiments allow a more controlled setting, reveal more detailed information about both participants, and - as a result - need smaller numbers of cases. Still, without going into the field, we will never confront our prototypes and design solutions with real pressure to maintain and facilitate the interviewer respondent interaction and the question answer process at the same time. Usability testing should therefore be seen as a joint process of laboratory experiments and field tests.

Acknowledgements

Some of these results were presented at the SMP Brown Bag Seminar, Institute for Social Research, University of Michigan on May 21, 1998 and on the occasion of the 54th Annual Meetings of the American Association for Public Opinion Research, May 16, 1999. Special thanks go to the interviewers participating in this experiment. Mick Couper, Siegfried Lamnek, Jeffrey Moore and two anonymous reviewers provided helpful suggestions on earlier versions of this paper.

References

Baker, R.P. (1992). New technology in survey research: computer-assisted personal interviewing (CAPI). *Social Science Computer Review*, 10, 145-157.

Couper, M.P., Baker, R.P., Bethlehem, J., Clark, C.Z.F., Martin, J., Nicholls, W.L. and O'reilly, J. (Eds.) (1998). *Computer Assisted Survey Information Collection*. New York: John Wiley & Sons, Inc.

Couper, M.P., and Burt, G. (1994). Interviewer attitudes toward computer-assisted personal interviewing (CAPI). *Social Science Computer Review*, 12, 38-54.

Couper, M.P., Groves, R.M. and Kosary, C. (1989). Methodological issues in CAPI. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 349-354.

Couper, M.P., Fuchs, M., Hansen, S.E. and Sparks, P. (1997). CAPI Instrument Design for the Consumer Expenditure (CE) Quarterly Interview Survey. Final Report. University of Michigan.

Fuchs, M. (1994) *Umfrageforschung mit Telefon und Computer*. Einführung in die computergestützte telefonische Befragung. Weinheim: Psychologie Verlags Union.

Fuchs, M. (1995). Die computergestützte telefonische Befragung. Einige Antworten auf Probleme der Umfrageforschung. *Zeitschrift für Soziologie*, 24, 284-299.

Fuchs, M. (2001). The impact of technology on interaction in computer-assisted interviews. (Ed. D.W. Maynard, H. Houtkoop-Steenstra, N.C. Schaeffer, and H. van der Zouwen). *Standardization and Tacit Knowledge*: *Interaction and Practice in the Survey Interview*. Wiley (forthcoming).

Fuchs, M., Couper, M. and Hansen, S. (2000). Technology effects: Do CAPI or PAPI interviews take longer? *Journal of Official Statistics* (in press).

Groves, R.M., and Mathiowetz, N.A. (1984). Computer assisted telephone interviewing: effect on interviewers and respondents. *Public Opinion Quarterly*, 48, 356-369.

Hansen, S.E., Couper, M.P. and Fuchs, M. (1998). Usability Evaluation of the NHIS Instrument. Paper presented at the Annual Meeting of the AAPOR, St. Louis, MO.

House, C.C. (1985). Questionnaire design with computer assisted telephone interviewing. *Journal of Official Statistics*, 1, 209-219.

House, C.C., and Nicholls, W.L. (1988). Questionnaire design for cati: design objectives and methods. (Ed. R.M. Groves, P.P. Biemer, L.E. Lyberg, J.T. Massey, W.L. Nicholls, and J. Waksberg). *Telephone Survey Methodology*, New York: Wiley, 421-436.

Laurie, H., and Moon, N. (1997). Converting to CAPI in a Longitudinal Panel Study. Working papers of the ESRC Research Centre on Micro-Social Change, 97-11, Essex.

Martin, E. (1999). Who knows who lives here? Within-household disagreements as a source of survey coverage error. *Public Opinion Quarterly*, 63, 220-236.

Moore, J.C. (1996). Person- vs. Topic-based Design for Computer-Assisted Household Survey Instruments. Paper presented at InterCASIC '96, International Conference on Computer-Assisted Survey Information Collection, San Antonio, TX.

Moore, J.C., and Moyer, H.L. (1998a). ACS/CATI Person-Based/Topic-Based Field Experiment - Final Report. Center for Survey Methods Research, U.S. Bureau of the Census.

Moore, J.C., and Moyer, H.L. (1998b). Questionnaire Design Effects on Interview Outcomes. Paper presented at the Annual Meeting of the AAPOR, St. Louis, MO.

Moyer, L.H. (1996). Which is better: grid listing or grouped questions design for data collection in establishment surveys? *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 986-990.

Nicholls, W.L., and De Leeuw, E. (1996). Factors in acceptance of computer-assisted interviewing methods: A conceptual and historic review. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 758-763.

Oksenberg, L., Beebe, T., Blixt, S. and Cannell, C. (1992). *Research on the Design and Conduct of the National Medical Expenditure Survey Interviews*. Final report. Survey Research Center, Ann Arbor, USA.

Oksenberg, L., Cannell, C. and Blixt, S. (1996). Analysis of Interviewer and Respondent Behavior in the Household Survey. U.S. Department of Health and Human Services. AHCPR No. 96-N016.

Projektgruppe Soep (1998). Funktion und Design einer Ergänzungsstichprobe für das Sozio-oeconomische Panel. Diskussionspapiere des DIW, 163, Berlin.

Schneid, M. (1991). Einsatz computergestutzter Befragungssysteme in der Bundesrepublik Deutschland. Ergebnisse einer Umfrage. ZUMA-Arbeitsbericht 91/20. Mannheim: ZUMA.

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

Suchman, L., and Jordan, B. (1990). Interactional troubles in face-to-face survey interviews. *Journal of the American Statistical Association*, 85, 45-54.

Weeks, M.F. (1992). Computer-assisted survey information collection: A review of CASIC methods and their implications for survey operations. *Journal of Official Statistics*, 8, 445-465.