Journal of Official Statistics, Vol. 26, No. 3, 2010, pp. 481–505

Comparison of Verbal Behaviors between Calendar and Standardized Conventional Questionnaires

Ipek Bilgen and Robert F. Belli[1]

This article examines verbal behaviors in 165 calendar and 162 standardized conventional interviews that collected life-course information on residence, cohabitation/marriage, children, education, labor and health from respondents in the Panel Study of Income Dynamics. A verbal behavior coding scheme was developed to categorize different interviewer and respondent behaviors used in the interviews. Using this coding scheme, we examined the differences in the usage of verbal behaviors between calendar and standardized conventional interviews. Findings revealed that calendar instruments encourage the more frequent use of (1) retrieval behaviors indicative of an attempt to encourage effective cuing of past events, (2) conversational behaviors that attempt to ensure a shared meaning between interviewers and respondents, and (3) interviewer probes that risk biasing respondents' answers. Although possible, it is unclear whether the flexible nature of calendar interviewing regarding the promotion of beneficial retrieval cues and conversational behaviors outweighs the increased risk of biasing respondents' answers.

Key words: Interviewer and respondent behaviors; retrospective reports; standardized interviews; conversational interviews; event history calendar.

1. Introduction

The collection of retrospective survey reports has been marked by two different interviewing and questionnaire design approaches. The most prevalent approach has been standardized interviewing, which has as its main goal the reduction of interviewer variance, which is mainly attempted by designing scripted questions and response options that are to be asked exactly as written by interviewers, and in which only nondirective probes are to be used (Beatty 1995; Fowler and Mangione 1990; Houtkoop-Steenstra 2000). This approach is so prevalent that it will be referred to as the conventional questionnaire (CQ) method throughout this article.[2] A less often used approach is Calendar interviewing,[3] which has as its main goal the optimizing of the accuracy of

[1] Both authors are at the University of Nebraska-Lincoln, Survey Research and Methodology (SRAM) program, 201 N 13th Street, Lincoln, NE 68588, U.S.A. Emails: ipekbilgen@huskers.unl.edu and bbelli2@unl.edu

[2] Conventional questionnaires have also been referred to as Question-list (Q-list) designs (Belli, Lee, Stafford, and Chou 2004; Belli, Shay, and Stafford 2001).

[3] In the literature, Calendar designs have been known by a variety of names, the most prevalent of these being the Life History Calendar and Event History Calendar.

Acknowledgments: Both authors contributed equally to this work, and their order was determined by a coin toss. This research has been supported by the National Institute on Aging, Grant number R01AG17977, to Robert Belli. The contents of this article are solely the responsibility of the authors and do not necessarily represent the official views of the National Institute on Aging or the National Institutes of Health. The authors especially thank Jamie Marincic, Yfke Ongena and Justin Wheeler for their contribution during coding scheme development and Debbie Miller for tracking and organizing the transcripts and coded documents.

© Statistics Sweden

autobiographical memory (Balán, Browning, Jelin, and Litzler 1969; Belli and Callegaro 2009; Freedman, Thornton, Camburn, Alwin, and Young-DeMarco 1988). In order to optimize autobiographical recall, Calendar methods encourage flexible interviewing in which interviewers are expected to use respondents' own past events as cues in the remembering of additional, temporally- and thematically-related events.

The goals of each of these approaches should be seen as ideals that can never be completely satisfied in practice. Examinations of the verbal behaviors of interviewers and respondents provide the ability to assess how well the goals of both CQ and Calendar methods have been met in practice. In her qualitative analysis of CQ interviews, Houtkoop-Steenstra (2000) found a number of verbal discrepancies from standardized interviewing ideals as interviewers will at times engage in directive behavior including negotiating answers with respondents. Some researchers have argued that standardization actually inhibits the benefits that natural conversation conveys in transmitting intended meanings across individuals (Schober and Conrad 1997; Suchman and Jordan 1990). Although Calendar interviewing has demonstrated that it usually provides better-quality retrospective data than does standardized interviewing (Belli, Shay, and Stafford 2001; Belli, Smith, Andreski, and Agrawal 2007; Engel, Keifer, and Zahm 2001; Yoshihama, Gillepsie, Hammock, Belli, and Tolman 2005), the recall that is elicited is far from perfect, and there are limitations that are imposed not only by errors that will populate any remembering task but also by poor interviewing practice. In addition to assessing the extent to which a method's specific goals are reached, another reason for examining verbal behaviors is that the context of retrospective reporting, in and of itself, will constrain the types of verbal behaviors that are generated by interviewers and their respondents regardless of method. Hence, those verbal behavioral differences that are observed between Calendar and CQ applications allow comparisons to be made of the potential advantages and disadvantages of using each respective method.

In an initial study examining the verbal behaviors of both Calendar and CQ interviews, Belli, Lee, Stafford, and Chou (2004) developed a comprehensive coding scheme to capture the linguistic expressions used by interviewers and respondents along several different dimensions. Because autobiographical recall is a key component of retrospective reporting, and because Calendar interviewing is designed to elicit effective recall cues, the types of retrieval probes provided by interviewers and retrieval strategies used by respondents were observed. Because respondent expressions of cognitive difficulty have constituted verbal behaviors that have been considered as useful in identifying poor question wording and challenging answering conditions (Belli and Lepkowski 1996; Fowler and Cannell 1996; Oksenberg, Cannell and Kalton 1991), these were also observed. Based on research that has emphasized the conversational character of survey interviews (Houtkoop-Steenstra 2000; Maynard and Schaeffer 2002), behaviors were identified that indicated attempts by interviewers and respondents to clarify meanings associated with specific question objectives or the survey process more generally. In standardized survey interviewing, nondirective interviewer feedback is encouraged to motivate respondents, and hence feedback, both nondirective and directive, was measured. Finally, rapport behaviors that are more directed toward interviewers and respondents dealing with each other as individuals engaged in a

social relationship, instead of targeting the survey questionnaire in and of itself, were identified and observed.

Belli (1998) hypothesized that, as compared to CQs, Calendar interviews optimize two types of recall cues, those associated with the parallel and sequential associations, among events stored in the structure of autobiographical memory. Parallel associations are drawn among events that occurred contemporaneously, or nearly so, such as a residential change that follows from a marriage. Sequential associations are drawn among events thematically similar as to which occurred earlier or later in time, such as remembering the order of the companies for which one has worked. With regard to sequential associations, Belli et al. (2004) observed four types among interviewer probes. Duration behaviors indicated a probe on the specific length of episodes, continuity behaviors indicated a concern for whether an episode continued to a specific date, timing was marked by an attempt to elicit from respondents the specific beginning or ending of episodes, and time gap fill is a behavior in which interviewers ask respondents to try to remember missing periods of time in a discontinuous timeline. As predicted, the occurrence of parallel behaviors occurred significantly more frequently in Calendar interviews than in CQ ones, both in terms of interviewers using parallel probes and respondents spontaneously engaging in parallel retrieval strategies. Not predicted was that parallel behaviors occurred rarely. The occurrence of sequential behaviors was more complicated. Whereas timing and time gap fill probes – and spontaneous sequential retrieval strategies – occurred more frequently in Calendar interviews, there was no difference between methods in the frequency of occurrence of continuity probes, and duration probes occurred more frequently in CQ interviews. This latter finding is affected by the design of the CQ questionnaire and the frequent presence of scripted questions asking respondents to report on the number of weeks episodes had lasted.

Not surprisingly, CQ interviews, as compared to Calendar ones, also led to more frequent use of probes that asked for behavioral frequencies (how many) and whether events had ever happened. These types of probes are not associated with drawing parallel and sequential associations, and asking how many times events happened is a common strategy used in CQ questionnaire design to determine behavioral frequencies (Belli and Callegaro 2009). In Calendar designs, behavioral frequencies are determined during the analytic phase on the basis of summing the reporting of specific events that are tied to specific points along respective timelines.

The cognitive difficulty behaviors included respondent requests for clarification, an attempt at a response that did not meet question objectives, the correction of earlier responses, and don't know responses, none of which differed between methods. Those behaviors that indicated conversational processes, interviewer verification, interviewer seeks clarification, and interviewer clarifies marked attempts to establish a shared meaning with respondents; and as one would expect, interviewer verification and seeking clarification occurred more frequently in Calendar interviews. Somewhat surprisingly, the prevalence of interviewer clarifies did not differ between methods.

Both feedback and rapport behaviors may assist in motivating respondents to remain interested in the questionnaire. In terms of feedback, it is deemed "acceptable" if it is non-affective and hence not likely to bias responses (e.g., a "thank-you"), and it is deemed "unacceptable" if it contains affect that may indicate to respondents approval or

disapproval by their interviewers. Task-related feedback occurs when interviewers refer to the task of the interviewing process, such as needing time to record responses. Both unacceptable and task-related feedback occurred more frequently in Calendar interviews, whereas the prevalence of acceptable feedback did not differ between conditions. Although the higher occurrence of unacceptable feedback in Calendar interviews indicates that there may be need for a greater concern about increased bias in Calendar interviews, observations of directive probing, which was assigned for any probe that was biasing, did not differ between conditions. The rapport behaviors of interviewer and respondent digressions and laughter also did not significantly differ in the frequency of occurrence between conditions.

Table 1 provides a summary of the findings of Belli et al. (2004). Taken together, these results confirm expected differences between Calendar and CQ interviews in the verbal behaviors that will be elicited during their administration. Calendar interviews elicit retrieval behaviors from both interviewers and respondents more often, which is consistent with their aim to maximize the recall of past events. In addition, the greater flexibility of Calendar interviews leads to more pronounced conversational and feedback behaviors, some of which raise concerns that interviewers may be biasing respondents in their answering. It is also notable that a number of behaviors, especially those associated with expressions of cognitive difficulty and rapport, did not differ between methods.

Although provocative, the results of Belli et al. (2004) are restricted in that only one set of interviews had been examined, in which both Calendar and CQ interviews were

Table 1. Summary of Belli et al. (2004) Findings

	More prevalent in:		
Category of Behavior	Calendar	CQ	No difference in prevalence
Retrieval	I parallel probes I timing probes I time-gap probes R parallel strategies R sequential strategies	I duration probes I how many probes I ever probes	I continuity probes
Cognitive difficulty			R requests for clarification R did not meet R correction R don't know
Conversational	I verifies I seeks clarification		I clarifies
Feedback	I unacceptable feedback I task-related feedback		I acceptable
Rapport			I digressions I laughter R digression R laughter

administered via phone using paper and pencil instruments, and for a reference period that asked respondents for events of the past two years. It is likely that certain behaviors will be affected by specific design considerations. We have seen that the frequent scripting of the question of "how many weeks" in this specific implementation of a CQ interview had likely impacted on the number of duration probes. In addition, the surprisingly low frequency of parallel behaviors in the Calendar interviews may have been impacted by using a relatively short (by Calendar standards) reference period of two years, which could have limited the opportunity for respondents to note associations among contemporaneous events. Finally, the specific manner in which behaviors are displayed is likely to be affected by mode and reference period considerations. Hence, the collection of additional verbal behavior data with another set of interviews can expand and clarify how Calendar and CQ interviews are differentially implemented.

In this article, we report on the results from the verbal behavior coding of Calendar and comparable CQ telephone interviews in which computer-assisted instruments were implemented and in which the reference period was the respondents' entire life course. The categories of verbal behaviors identified and measured were similar to those investigated by Belli et al. (2004), though because of differences in method and reference period they were not identical. Our expectations were that the same pattern of results would emerge with more retrieval and flexible conversational behaviors being elicited in the Calendar interviews, and that CQ design considerations in terms of the types of questions that were asked to collect retrospective reports would increase the prevalence of certain behaviors. Finally, we expected that there would be a greater prevalence of parallel behaviors as asking questions about the life course provides respondents with more opportunities to observe contemporaneous events.

2. Methods

2.1. Data Collection

At the first stage of the study, a nationwide subsample of 632 participants from the 2001 Panel Study of Income Dynamics (PSID)[4] were randomly selected and interviewed via telephone from July through September 2002. Participants in this experiment included only the members of PSID households (family units) who had participated every year (in every wave of PSID) at least from 1980 to 1997 and had been interviewed during at least half of the waves in which their families participated in the study. In this study, 632 respondents and 26 interviewers were randomly assigned either to a Calendar condition ($N_{respondent} = 313$, $N_{interviewer} = 13$) or to a Standardized CQ condition ($N_{respondent} = 319$, $N_{interviewer} = 13$). Both Calendar and CQ participants were offered a $50 incentive for their participation and all respondents were interviewed with Computer Assisted Telephone Interviewing (CATI). The CQ condition had been programmed within the prepackaged *Blaise* ® language and used a traditional computer assisted standardized

[4] PSID is a longitudinal study of a probability sample of the U.S. households (family units) which involved interviewing and reinterviewing members of sampled families – whether or not they were living in the same dwelling or with the same people – every year from 1968 to 1997. PSID followed members of the households as they grew older, and as they formed family units of their own.

survey instrument, whereas the Calendar condition had been programmed in-house at the University of Michigan. With the permission of the respondents, 93% of the interviews ($N_{CAL} = 297$; $N_{CQ} = 291$) were tape recorded. Of the interviews that were tape recorded, 4% were found to be problematic either due to poor sound quality or case identification mismatch between the audio-tape and the response data file. All available nonproblematic tapes were then transcribed ($N_{CAL} = 291$; $N_{CQ} = 273$).

Interviews in both methods were designed to asked respondents to report on their life course experiences (from birth until the time of the interview) with residential changes, marriages, cohabitations, and the persons whom they had parented, their labor histories (especially employment and unemployment), and certain aspects of their health histories (disability, health status, smoking, and weight). Respondents were also asked about their formal relationships to the persons who had raised them, and their socioeconomic status while growing up. There were 26 interviewers, who were first matched on the basis of interviewing experience and then randomly assigned to one of the methods; this assignment procedure led to interviewers who had equivalent experience between methods not only in prior experience in both telephone and face-to-face interviewing, but also when it came to gender (3 and 2 male Calendar and CQ interviewers, respectively) and age in years.[5]

2.2. *Verbal Behavior Coding*

Using the verbal behavior codes developed by Belli et al. (2004) as a guide, a new verbal behavior coding scheme was developed through research group discussions that included as members both authors, a Ph.D. level consultant, three graduate students, and an undergraduate student with transcribing experience who later also served as a coder. Coder training documents were created, coders were trained and the verbal behavior coding scheme was developed and improved in 70 meetings which on average lasted approximately an hour and a half each. It was established at the start of this process that a turn would be considered the unit for coding. A turn is defined as a continuous stream of speech undertaken by either the interviewer or the respondent, as perceived by the transcribers in their transcriptions. At the first 35 meetings, the research group developed the core coding scheme and coded Calendar and CQ interviewer scripts and training documents. In the remaining thirty-five meetings, a total of five undergraduate students, who had prior transcribing experience, were trained as coders with the coding scheme being improved concurrently with the identification of additional behaviors to identify in the scheme. In the training meetings, various problematic behaviors were discussed and new rules and solutions were developed by discussion and consensus of the members of the coding group. Via these discussions, consistency among coders was enhanced to the point where there was confidence that reliability would be maintained in the completely independent production coding.

At the initiation of production coding, Calendar and CQ transcripts were randomly divided evenly into three replicates, with each of the five coders randomly assigned to an even portion of the tapes in each replicate and for each condition. Coders had available to

[5] Belli, Smith, Andreski, and Agrawal (2007) provides greater detail of the study's methods.

them a finalized coding scheme which incorporated 30 interviewer and 29 respondent verbal behavior codes. Coders were also provided with pre-coded CQ and Calendar scripts from which they could decide appropriate codes. For each replicate, a randomly selected 10% of the transcripts were double coded by a master coder (one of the graduate students who had been involved in the research group) in order to assess inter-coder reliability. Each replicate was finished completely before any coder was allowed to move on to the next replicate. Due to the budget constraints, 55% of the recorded tapes (i.e., entire first replicate and 69% of the second replicate) were coded. As a result, in this article we will focus on 327 coded transcripts (165 Calendar and 162 CQ) in order to discuss verbal behavior comparisons between Calendar and CQ interviewing methods.

2.3. Coding Scheme

The coding scheme was designed to capture several different dimensions of the verbal behaviors used by interviewers and respondents. Identified are interviewer behaviors including the use of retrieval probes, standardization associated behaviors, conversational behaviors, feedback, and rapport behaviors. Table 2 provides categories of these interviewer behaviors, behavior definitions and examples. Furthermore, the coding scheme captures respondents' use of retrieval strategies, responses to targeted probes, cognitive difficulty related behaviors, as well as conversational and rapport behaviors. Table 3 provides categories of these respondent behaviors, behavior definitions and examples.

As autobiographical recall is a key component of retrospective reporting, included are behaviors that measure the occurrence of interviewer retrieval probes. Three types of probes (parallel, sequential and data element) reflect the associations among events stored in the structure of autobiographical memory. *Parallel retrieval* probes associate contemporaneous events from different domains. Parallel probes also include reference to a *holiday* or to a landmark *historical* event. *Sequential retrieval* probes associate thematically similar events as to which occur earlier or later in time. Sequential retrieval probes include *duration* probes (which seek the specific length of episodes, periods of time, or what will be referred to in the remainder of this article as *spells*), *timing* probes (which seek the specific beginning or ending of spells), and undifferentiated *sequential* probes of what spells happened beforehand or next. *Data element* probes associate specific events with more abstract or general spells that have already been identified by the respondent. Moreover, the coding scheme also includes time location probes such as interviewers' probing with *age*, *season*, or *year*, and other retrieval probes such as whether a type of spell *ever* occurred; or *how many* times an event occurred.

Consistently for respondents, the coding scheme is designed to measure the occurrence of three types of retrieval strategies (respondent parallel, sequential and data element behaviors) that reflect respondents' spontaneous associations among events stored in the structure of autobiographical memory. Respondent *parallel retrieval* strategies spontaneously associate contemporaneous events from different life domains. Parallel retrieval strategies also include spontaneous references to a *holiday* or to an *historical* event. Respondents' *sequential retrievals* exist as spontaneous associations of events thematically similar as to which occurred earlier or later in time. The sequential strategies

Table 2. Interviewer Verbal Behaviors and Descriptions (definitions and examples)

Verbal behaviors	Descriptions
(1) Retrieval probes	
a. Parallel retrieval probe	
Parallel	Interviewer uses an event from the respondent's past as an anchor. This event is not part of the domain being administered. Example: *When you got married. . .that would be in May then of. . . (used in residence domain)*
Holiday	Interviewer uses a public holiday as a parallel probe. Example: *. . .can you remember kind of the Christmas season at the gift shop?*
Historical	Interviewer uses a historical landmark event as a parallel probe. Example: *Do you remember if that was before or after John Kennedy was shot?*
b. Sequential retrieval probes	
Duration	Interviewer is seeking how long a spell has occurred. Example: *How long did you work for them?*
Timing	Interviewer is seeking when a spell began or ended. Example: *And please tell me again when you were married.*
Sequential	Interviewer is probing for data elements of a spell that happened earlier or later and has not yet been explicitly temporally defined in any way. Example: *Okay, uh, sir, can you tell me where you lived before you moved to city1?*
c. Top-down retrieval probe	
Data elements	Interviewer is specifically seeking data elements as part of a spell that has already been temporally defined in some way. Example 1: *May I have that address?*
d. Time location probes	
Interviewer uses age	Interviewer is using age information to keep track of a spell. Example: *When did you move out of your home when you were 22, do you remember what month it was?*
Interviewer uses season	Interviewer is using season information in some spell related way. Example: *Do you remember if it was winter, or. . .?*
Interviewer uses year	Interviewer is using year information to keep track of a spell. Example: *And then from May of '84 until. . .?*
e. Other retrieval probes	
Ever	Interviewer is probing for whether a type of spell (which is listed in the "Types of Spells and Data elements document") ever happened. Example: *Have you ever lived with a partner as if married?*

Table 2. Continued

Verbal behaviors	Descriptions
How many	Interviewer asks about the number of spells, data elements, or any non-temporal entities. Example: *How many jobs do you currently have?*
(2) Standardization associated behaviours	
Significant change	A significant change in question wording is one that appears to change, or can conceivably change, the meaning of a scripted question. A potential change in the meaning is a "Significant Change" hence two or more adjacent substantive words (words other than "a", "the", "or", etc. . .), or any phrase that is added or missing from the script, is coded as significant change.
Directive	Interviewer provides any probe that poses the risk of biasing the respondent's answer. Example: *Do you remember if it was winter, or. . .? (In this case, a nondirective way of asking the question can be "do you remember which season this incident occurred?")*
Misstate	Interviewer probes with incorrect information. Misstate (M) can be double-coded. Example: *Alright. So then from January of '96 to. . . (correct year: '76)*
Question repair	Interviewer repairs a question or part of a question, if there is a significant change earlier. Example: *For 3 months or more, did you work more than 3 month – 3 months or more in the summer?*
Question repeat	Interviewer repeats all or part of a scripted question, if there is no significant change earlier. Question Repeat is coded when the scripted question is repeated. Example: *So, let me go back again. Thinking over your entire life, from birth to the present, have you ever been hospitalized for one month or more? [underlined is repeated portion]*
Scripted directive	Interviewer asks a question that is literally provided by the script which may bias the respondent's answer. Example: *Are you still living with this partner?*
(3) Conversational behaviors	
a. Uncertainty behaviors	
Interviewer seeks clarification	Interviewer seeks clarification from respondent on some aspect of the survey/questionnaire that can impact satisfying the questionnaire/survey objectives. The information sought is beyond what has already been given or is a request, by the interviewer to the respondent, to repeat what has already been given. Example: *Respondent: During the summers – I spent summers in State 2, does that make a difference? Interviewer: Uh, you very – like on a vacation there?*

Table 2. Continued

Verbal behaviors	Descriptions
Interviewer verification	Interviewer uses information that has already been provided by the respondent to either confirm the accuracy of the information or allow the interviewer to convey to the respondent what information they are currently focusing on. Example: *You lived in the same street again, at a different address, is that true?*
Interviewer spells out	Interviewer spells out a word. Example: *And I will just verify that with you. Ok, A-d-d-r-e-s-s-1?*
b. Response to perceived cognitive problems	
Interviewer gives clarification	Interviewer provides *unscripted clarification* on some aspect of the survey/questionnaire. Example: *R: Let me ask you this. Now you said technical I went to x-ray training I was, uh, out of high school. I: I believe that would be considered, um, vocational school.*
Scripted interviewer clarification	Interviewer provides scripted clarification of what to include/not include, or an explanation of how to go about answering the questionnaire. Example: *If you did not graduate from high school, but took classes to earn a GED, I would like to know about this as well.*
(4) Feedback behaviors	
Task related feedback	Interviewer refers to some logistical (operational) task-related component of the interviewing process (i.e., instrument/questionnaire). Example: *I will go back here and we will correct this one.*
Acceptable feedback	A neutral phrase following a response to a study-relevant probe that shows appreciation for receipt of the response. Example: *Thank you.*
Unacceptable feedback	Nonneutral phrases following a response to a study-relevant probe. Example: *That sounds good.*
Interviewer redirects	Interviewer redirects the respondent to another question. Example: *Ok, well we'll get into that after–*
(5) Rapport behaviors	
Interviewer digression	Interviewer asks a question or makes a comment that is not a direct attempt to satisfy study or questionnaire objectives. Example: *Oh, hang on to those, that will be invaluable later on for your family.*
Interviewer distancing	Interviewer makes a comment that provides information to the respondent that the questions originate with a third party, the survey researcher. Example: *I have to ask you the questions; I mean I'm not trying to be repetitious.*
Interviewer laughs	Interviewer laughs during the interview.

Table 3. Respondent Verbal Behaviors and Descriptions (definitions and examples)

Verbal behaviors	Descriptions
(1) Retrieval strategies	
a. Parallel retrieval	
Respondent parallel	Respondent spontaneously refers to a contemporaneous state or event in an area different from the required elements of a domain. Example: *It was football season when it started up.*
Respondent holiday	Respondent spontaneously refers to a holiday. Example: *She was supposed to be born on a holiday too. Oh, around Christmas time.*
Respondent historical	Respondent spontaneously refers to a historical landmark event. Example: *That'd be 1930 up to 1939 were the Depression years really. . .*
b. Sequential retrieval	
Duration response	Respondent spontaneously provides how long a spell occurred. Example: *I worked for a year.*
Timing response	Respondent spontaneously provides the beginning or ending of a spell, or spontaneously indicates any specific date. Example: *Um, so that would have been September of the year prior.*
Sequential response	Respondent spontaneously provides a data element for a spell that occurred earlier or later and has not yet been explicitly temporally defined in any way. Example: *So, if three months was a summer job, I guess it doesn't count.*
c. Top-down retrieval	
Data Element Response	Respondent spontaneously provides a data element for a spell which is already temporally defined in some way. Example: *Until May of '85, I moved back to address1.*
d. Time location strategies	
Respondent uses age	Respondent uses an age to define the beginning or ending of a spell. Example: *I started my first job when I was 15.*
Respondent uses season	Respondent uses a season to define the beginning or ending of a spell. Example: *In the winter, um. . . (Respondent is using season in order to remember when he started his first job)*
Respondent uses year	Respondent uses a year to define the beginning or ending of a spell. Example: *In '79, let's see. Until 1983.*
(2) Responses to targeted probes	
a. Agreement	Respondent agrees with interviewer's probe
Directive agreement	Respondent agrees with an interviewer's directive probe. Example: *Interviewer: Near city1. Respondent: Yes, near city1, you are right.*

Table 3. Continued

Verbal behaviors	Descriptions
Verification agreement	Respondent confirms interviewer's verification. Example: *Interviewer: You joined in '61 and you stayed in for 20 years. Respondent: Yes ma'am.*
Misstate agreement	Respondent agrees with the interviewer's misstatement. Example: *Interviewer: So, you moved out March of '73 (Interviewer means 1993). Respondent: Yes.*
b. Disagreement	Respondent disagrees with interviewer's probe
Directive disagreement	Respondent disagrees with interviewer's directive probe. Example: *Interviewer: Okay, but do you remember if it was winter, or. . .? Respondent: It was spring.*
Verification disagreement	Respondent disagrees with interviewer's verification. Example: *Interviewer: And, now did you say State2? Respondent: State 1*
Misstate disagreement	Respondent disagrees with interviewer's misstate. Example: *Interviewer: So when did you get out in 1961, what month–, (interviewer means 1981). Respondent: I didn't get out.*
c. Undetermined	It cannot be determined if the respondent agrees or disagrees with interviewer's probe.
Directive undetermined	It cannot be determined if the respondent agrees or disagrees with interviewer's directive probe. Example: *Interviewer: Until the present? Respondent:: I feel a little overweight, but not like I used to be.*
Verification undetermined	It cannot be determined if the respondent agrees or disagrees with interviewer's verification. Example: *Interviewer: And that was sta–; Respondent: State 2. (We can not tell if the respondent agrees or disagrees about the state).*
Misstate undetermined	It cannot be determined if the respondent agrees or disagrees with interviewer's misstate. Example: *Interviewer: . . .and then in May of '84? Respondent: I lived um–until May of '85 at City1. . .*
Nothing new	Respondent informs interviewer that there is no new spell or new data element to enter into survey instrument (this includes "no" responses to an ever probe). Example 1: *I: Did you ever graduate from college? R: No, I did not.* Example 2: *I: For three months or more, did you work more than three month – three months or more in summer? R: No.*
(3) Cognitive difficulty behaviors	
Request for clarification	The respondent indicates that more information is needed about the study or to answer a question. This includes requests that a question be repeated. Example: *Interviewer: And then you moved out of there in June? Respondent: Where's that?*

Table 3. Continued

Verbal behaviors	Descriptions
Qualified response	A phrase that causes a response to be qualified in some way, or made to seem less than certain (e.g., probably, I guess, I think, depends, around, about, maybe, not really sure, but, I would say). Example: *I guess it was January.*
Does not meet	Respondent attempts to answer the question but fails to meet questionnaire objectives including not completely verbalize a response and providing multiple possible answers without choosing a single one. Example: *Interviewer: In what month and year did you start with employer1 as your main job? Respondent: Oh, dear, couple a–worked for him of–couple of times. Maybe July or August.*
Correction	Respondent corrects an earlier substantive response.
Don't know	Respondent indicates that he/she does not know the answer of the question which is asked by the interviewer. Example: *No ma'am, I don't know the months, I can't remember the exact dates on that.*
(4) Conversational behaviours	
Respondent offers or provides clarification	Respondent offers to provide clarification or provides clarification on any aspect of study-relevant information, either spontaneously or when the interviewer requests clarification. Example: *Well, I could give you the jobs, but I sure couldn't give you the dates.*
Respondent spells out	Respondent spells out a word.
(5) Rapport behaviors	
Respondent digression	Respondent makes a comment that is not a direct attempt to satisfy study or questionnaire objectives. In other words, Respondent Digression is not an attempt at a response. Example: *Those beeps you are hearing are somebody calling in, and they can, they can call back later.*
Respondent laughs	Respondent laughs.

include behaviors in which respondents spontaneously provide how long a spell occurred (*duration* responses), specific reference to the ending or beginning of spells (*timing* responses), and undifferentiated *sequential* responses in which events that happened earlier or later are spontaneously reported. Also, the coding scheme permits measurement of *data element* responses in which respondents spontaneously associate specific events with more abstract periods of time or spells that have already been identified. With the aim of exploring interviewer and respondent retrieval behaviors, the coding scheme takes into consideration the usage of time location strategies in which respondents spontaneously use *age*, *season* or *year* in the remembering of past events.

Another purpose of this study is to investigate interviewer behaviors that can potentially bias respondents' answers. As a result, we coded for *significant* changes in question wordings as well as *directive* probes, or probes which are included in the script but still might bias respondents' answers (*scripted directive*). Also, transcripts revealed instances in which interviewers either probed respondents with incorrect information (*misstate*); repaired a question or a part of question if there had been a significant change earlier (*question repair*); or repeated a scripted question if there had not been a significant change earlier (*question repeat*). We also measured interviewer attempts to establish a shared meaning with respondents such as interviewers' attempts to confirm the accuracy of answers (*interviewer verification*) or to clarify the answers which respondents had provided earlier (*interviewer seeks clarification*). Our coding scheme also includes respondents' *agreements* and *disagreements* with interviewers' *directive, verification* and *misstate* probes, included as types of behaviors in which there were responses to interviewer queries that targeted specific responses. If it cannot be determined whether respondents agree or disagree with interviewers' *directive, verification* and *misstate* probes, the coding scheme provides observation of *undetermined* behaviors (*directive undetermined, verification undetermined* and *misstate undetermined*). Additional responses to targeted probes include a statement by respondents that there is nothing to add to what has already been reported (*nothing new*), and when respondents choose to respond to either a duration or timing probe with one type of answer, but not both (*choice*).

Respondent cognitive difficulty behaviors such as indications regarding when more information is needed to answer a question (*request for clarification)*, uncertainty phrases which qualify a response (*qualified response*), *don't know* responses, *corrections* of an earlier response, and attempts to answer a question which fail to meet the questionnaire objectives (*does not meet*) were also observed and included in the coding scheme. As a result, interviewer responses to these perceived cognitive problems in order to resolve uncertainties were also observed. These interviewer behaviors were coded as *interviewer gives clarification* and *scripted interviewer clarification*. In addition, respondents' spontaneous attempts to *offer or provide clarification* on any aspects of the study-relevant information, and interviewers' and respondents' attempts to *spell out* words are considered as conversational behaviors in the coding scheme.

Finally, the coding scheme captures interviewer feedback and rapport behaviors used in order to assist in motivating respondents to remain interested in the questionnaire. The type of interviewer feedback that is not likely to bias responses (e.g., a "thank-you") is considered as *acceptable feedback*; whereas feedback that may indicate interviewers' approval or disapproval in their answers is coded as *unacceptable feedback*. In addition,

task-related feedback behaviors received a verbal behavior code when interviewers referred to the task of the interviewing process, such as needing time to record responses. The coding scheme also considers the *interviewer redirecting* behaviors which occurred when interviewers redirected respondents to another question. Lastly, interviewer and respondent rapport behaviors that appear to be more directed toward attempts to build interpersonal communication (instead of targeting the survey questionnaire objectives) were observed and identified in the coding scheme. Interviewer rapport behaviors include interviewers' comments or questions that digress from the study or questionnaire objectives (*interviewer digressions)*, interviewers' comments that provide information to the respondent that the questions originate with the researcher (*interviewer distancing*), and *interviewer laughter*. Respondent rapport behaviors include respondents' comments or questions that digress from the study or questionnaire objectives (*respondent digressions*) and *respondent laughter*.

3. Results

3.1. Inter-coder Reliabilities

Out of 327 coded transcripts, 9.79% randomly selected transcripts (18 Calendar and 14 CQ) were double coded by a master coder. The master coder has been in the project from the beginning of the coding scheme development process. The purpose of the double-coding was to examine the inter-coder reliabilities as measured by Kappa for each coded behavior and to exclude the verbal behavior codes which are not reliability coded. A kappa value equal to or greater than .40 indicates an adequate level of agreement (Bartko 1966; Cohen 1960; Fleiss 1971) and has been used as the traditional standard criterion of reliability for behavior coding (Oksenberg, Cannell, and Kalton 1991; Presser and Blair 1994). Agreements and disagreements as to whether each behavior occurred were tallied for each turn of the double coded interviews, and these tallies were the basis of the resulting kappa indices for each behavior as reported in Table 4.

Inter-coder reliabilities are adequate (kappa values are higher than 0.40) for 27 interviewer verbal behaviors, but are below 0.40 for interviewer verbal behaviors such as *distancing, significant change* and *question repair*. Table 5 provides kappa calculations for each respondent behavior in which the turns are the unit of analyses. Inter-coder reliabilities are adequate for 24 respondent verbal behaviors, but are below 0.40 for respondent verbal behaviors including *misstate agreement*, *verification disagreement,* and each of the *undetermined* behaviors following a *directive, verification,* and *misstatement,* respectively. Codes that have failed to reach a kappa of .40 are excluded from further analysis due to a nonadequate level of agreement.

3.2. Calendar and CQ Verbal Behavior Comparisons

Between the 165 Calendar and 162 CQ interviews, independent *t*-tests were computed on the mean total occurrences of each verbal behavior per interview. As multiple comparisons were conducted, we controlled for Type I errors, at $\alpha = .05$, by adjusting the *p*-values separately for interviewer and respondent behaviors using a

Table 4. Calendar and CQ Kappa and t-test Results for the Interviewer Verbal Behaviors

Interviewer verbal behaviors	Kappa values	Calendar		CQ		Mean differences (SEs)
		Mean	SD	Mean	SD	
(1) Retrieval probes						
a. Parallel retrieval probe						
Parallel	0.527	3.661	5.580	0.525	1.196	3.136 (0.44)***
Holiday	0.667	0.121	0.363	0.006	0.079	0.115 (0.03)***
Historical	0.667	0.115	0.419	0.006	0.079	0.109 (0.03)***
b. Sequential retrieval probes						
Duration	0.852	11.70	6.81[a]	1.59	1.95[a]	10.12 (0.55)***
Timing	0.862	44.01	15.13[a]	47.63	19.2[a]	−3.62 (1.91)
Sequential	0.758	7.95	6.35	7.55	6.35	0.40 (0.70)
c. Top-down retrieval probe						
Data elements	0.780	33.73	12.57[a]	48.05	17.75[a]	−14.32 (1.70)***
d. Time location p						
Interviewer uses age	0.741	8.85	6.17[a]	7.25	4.77[a]	1.60 (0.31)*
Interviewer uses season	0.600	1.61	2.18[a]	0.23	0.69[a]	1.37 (0.18)***
Interviewer uses year	0.853	44.58	25.40	47.31	27.38	−2.73 (2.92)
e. Other retrieval probes						
Ever	0.906	17.74	5.19[a]	48.52	18.16[a]	−30.78 (1.48)***
How many	0.901	0.87	1.38[a]	4.72	2.66[a]	−3.84 (0.23)***
(2) Standardization associated behaviors						
Directive	0.658	31.27	13.10[a]	8.20	7.96[a]	23.08 (1.20)***
Misstate	0.479	0.90	1.33[a]	0.27	0.58[a]	0.63 (0.11)***
Question repeat	0.497	1.28	1.90[a]	5.92	7.15[a]	−4.63 (0.58)***
Scripted directive	0.745	0.71	1.49[a]	13.92	5.88[a]	−13.21 (0.48)***
(3) Conversational behaviours						
a. Uncertainty behaviors						
Interviewer seeks clarification	0.466	11.93	7.95	10.81	7.99	1.11 (0.88)
Interviewer verification	0.686	34.76	21.38[a]	22.76	16.09[a]	12.00 (2.09)***
Interviewer spells out	0.886	2.49	3.32[a]	4.59	6.64[a]	−2.10 (0.58)***

Table 4. Continued

Interviewer verbal behaviors	Kappa values	Calendar		CQ		Mean differences (SEs)
		Mean	SD	Mean	SD	
b. Response to perceived cognitive problems						
Interviewer gives clarification	0.567	18.23	12.76	13.42	11.12	4.81 (1.32)***
Scripted interviewer clarification	0.739	7.71	5.02[a]	34.67	15.37[a]	−26.96 (1.27)***
(4) Feedback behaviors						
Task related feedback	0.726	9.37	9.59[a]	6.64	5.78[a]	2.73 (0.87)***
Acceptable feedback	0.925	6.31	8.28[a]	16.65	15.76[a]	−10.34 (1.40)***
Unacceptable feedback	0.694	6.34	6.74	4.57	6.38	1.77 (0.73)*
Interviewer redirects	0.476	0.51	0.93[a]	0.37	0.76[a]	0.14 (0.09)
(5) Rapport behaviors						
Interviewer digression	0.588	7.19	8.27[a]	3.89	6.26[a]	3.30 (0.81)***
Interviewer laughs	0.856	11.93	12.38[a]	15.64	14.51[a]	−3.71 (1.49)*

* $p < .05$; ** $p < .01$; *** $p < .005$.

[a] Equal variances assumption does not hold (the variances across two groups are heterogeneous).

Table 5. Calendar and CQ Kappa and t-test Results for the Respondent Verbal Behaviors

Respondent verbal behaviors	Kappa values	Calendar		CQ		Mean differences (SEs)
		Mean	SD	Mean	SD	
(1) Retrieval strategies						
a. Parallel retrieval						
R Parallel	0.730	11.273	10.714	7.519	9.030	3.754 (1.10)***
R Holiday	0.588	0.370	0.970	0.167	0.526	0.203 (0.09)*
R Historical	0.625	0.358	1.000	0.327	1.387	0.030 (0.13)
b. Sequential retrieval						
Duration response	0.750	7.27	5.30[a]	5.23	6.42[a]	2.04 (0.65)***
Timing response	0.625	15.69	12.00[a]	7.68	7.64[a]	8.01 (1.11)***
Sequential response	0.541	4.67	5.31	5.06	5.46	−0.40 (0.60)
c. Top-down retrieval						
Data element response	0.419	12.55	9.15	10.99	9.02	1.56 (1.01)
d. Time location strategies						
Respondent uses age	0.876	6.74	4.64	4.85	5.28	1.89 (0.55)***
Respondent uses season	0.824	3.78	4.63[a]	1.93	2.75[a]	1.84 (0.42)***
Respondent uses year	0.856	31.68	18.24	37.72	18.18	−6.04 (2.01)***
(2) Responses to targeted probes						
Directive agreement	0.680	21.497	9.139	13.710	7.533	7.787 (0.93)***
Verification agreement	0.696	27.467	17.256	17.179	11.565	10.288 (1.62)***
Directive disagreement	0.530	4.346	2.711	4.975	2.507	−0.630 (0.29)*
Misstate disagreement	0.512	0.539	0.940	0.179	0.497	0.360 (0.83)***
Nothing new	0.894	8.18	3.10[a]	20.73	7.77[a]	−12.55 (0.66)***
(3) Cognitive difficulty behaviors						
Request for clarification	0.906	8.56	7.02[a]	9.63	9.12[a]	−1.07 (0.90)
Qualified response	0.727	17.79	12.94	17.28	11.24	0.52 (1.34)
Does not meet	0.405	5.07	4.82	5.16	5.53	−0.09 (0.57)
Correction	0.674	2.07	2.23[a]	1.88	1.90[a]	0.19 (0.23)
Don’t know	0.764	7.48	6.82[a]	11.86	11.44[a]	−4.37 (1.04)***

Table 5. Continued

Respondent verbal behaviors	Kappa values	Calendar		CQ		Mean differences (SEs)
		Mean	SD	Mean	SD	
(4) Conversational behaviors						
Respondent offers or provides clarification	0.551	39.54	24.25	34.49	24.10	5.05 (2.67)
Respondent spells out	0.915	4.96	5.06[a]	6.28	6.97[a]	−1.31 (0.67)
(5) Rapport behaviors						
Respondent digression	0.471	8.23	10.67[a]	5.08	8.68[a]	3.15 (1.07)***
Respondent laughs	0.890	11.53	14.53[a]	11.81	12.22[a]	−0.28 (1.49)

* $p < .05$; ** $p < .01$; *** $p < .005$.

[a] Equal variances assumption does not hold (the variances across two groups are heterogeneous).

Benjamini-Hochberg test. Table 4 provides the mean number of occurrences of each interviewer verbal behavior in both Calendar and CQ interviews, and also includes Calendar and CQ interviewer verbal behavior mean differences and significance of the mean differences using the independent *t*-tests. The same types of statistics are provided for respondent behaviors in Table 5.

Interviewers are significantly more likely to use retrieval probes such as *parallel retrieval (parallel, holiday, historical), duration, interviewer uses age* and *interviewer uses season* in the Calendar interviews than in the CQ interviews, whereas the *data elements, ever,* and *how many* retrieval probes occurred significantly more often in the CQ interviews than in the Calendar ones. These findings generally support the expectation that Calendar interviews encourage interviewers to use parallel and sequential probes, whereas CQ interviews encourage the use of probing strategies that are not as sensitive to the structure of autobiographical memory (Belli 1998; Belli and Callegaro 2009; Freedman et al. 1988).

As for the retrieval strategies used by respondents (see Table 5), they also reflect the significantly greater use of spontaneous parallel and sequential strategies in Calendar interviews in than CQ ones, as seen in the *respondent parallel*, *holiday*, *duration, and timing behaviors.* Notably, the occurrence of *sequential* retrieval strategies did not significantly differ between methods, perhaps because in CQ interviews (as in Calendar ones) respondents became knowledgeable about what would likely be asked next. *Respondent uses age and season* are more prevalent in Calendar interviews, whereas in CQ interviews respondents are more likely to spontaneously use *year* information than in Calendar ones. It should be noted that the use of *age* and year provides redundant information, and that the greater use of *age* in Calendar interviews by both interviewers and respondents indicates greater flexibility in determining the location of events in time, as the default for CQ interviews had been to use *year* despite a design that provided a year to age conversion in these interviews. The greater use of *season* by both interviewers and respondents in Calendar interviews indicates a heightened awareness that contemporaneous context is helpful in remembering and locating events.

Taken together, several of the interviewer standardization associated behaviors and feedback behaviors illustrate potential instances in which interviewers run the risk of biasing the respondents' answers (Cannell, Miller, and Oksenberg 1981), which is more pronounced in Calendar interviewing due its more flexible nature. Interviewers are more likely to offer *directive* probes, to probe with incorrect information (*misstate behaviors*), and to provide *unacceptable feedback* in Calendar interviews. In contrast, CQ interviewers more often provide *acceptable feedback*. The greater flexibility in Calendar interviews also leads to the more frequent use of conversational behaviors among interviewers including *verification* and *giving clarification,* and is likely implicated in the more frequent occurrence of *task related feedback*, Regarding task related feedback, interviewers in Calendar designs will need to make fairly frequent adjustments in entering information deriving from respondents' offering responses in a less constrained manner than is typically observed in CQ interviews.

Although Calendar interviews appear to encourage the more frequent occurrence of potentially biasing interviewer behaviors in comparison to CQ instruments, it is uncertain whether data quality is in fact adversely impacted by these behaviors. Comparisons in data

quality between Calendar and CQ designs consistently find more advantages with the Calendar (Belli, Shay, and Stafford 2001; Belli et al. 2007; Engel, Keifer, and Zahm 2001; Yoshihama et al. 2005). One possibility that needs further research is that any biasing probing may be offset by the frequent use of conversational behaviors in Calendar interviews that are beneficial to data quality, including interviewers verifying respondents' answers and assisting respondents in understanding the nature of the questions by providing clarification. It is noteworthy that certain conversational behaviors may be typically scripted in the questions that are written for CQ instruments, as evidenced by the frequent occurrence of *scripted interviewer clarification* in the current results for the CQ. In addition, CQ interviews also might promote some level of biasing probing. Although interviewers in CQ instruments are less likely to spontaneously probe directively, it may be fairly common to write directive scripts in CQ questions, as had occurred with the current CQ instrument, as revealed by the substantive prevalence of *scripted directive* queries.

As for the respondents, they are in more frequent agreement with interviewers' directive probes and verifications in the Calendar interviews than in the CQ ones as a function of these interviewer behaviors occurring more frequently in Calendar interviews. Agreements verifications are likely indications that interviewers had understood respondents correctly. As for agreements directive probing, there is considerable ambiguity in interpretation because although they may represent an impact of biasing, it may also be the case that interviewers were usually correct in the assumptions that had led to their framing of these directive probes (see, for example, Stafford and Belli 2009, Figure 17.1)[6].

There is indirect evidence favoring the latter interpretation. Mirroring the more frequent occurrence of misstatements in Calendar interviews, respondents more frequently disagree with interviewers' misstatements in Calendar interviews in comparison to those in the CQ. Hence, respondents appeared to have been able to mostly catch any misstatements that were presented to them. Assuming that respondents would be equally sensitive to catching directive behaviors as they were to catching misstatements that were not reflective of their circumstances, it is noteworthy that respondents disagreed more frequently with interviewer directive behaviors in CQ interviews in than Calendar ones, despite directives occurring more frequently with Calendar interviewers. This pattern of data suggests that interviewer directives in Calendar interviews were based on correct inferences more often than when directives appeared among CQ interviews.

Moreover, there was little difference between methods in the frequency of occurrence of respondent cognitive difficulty behaviors, which are behaviors that are most suggestive of problems associated with poorer data quality (see also Fowler and Cannell 1996). If anything, CQ interviews were more problematic, as indicated by the more frequent occurrence of *don't know* statements by respondents.

Rapport behaviors are also not clearly diagnostic of data quality, although some researchers have suggested that rapport can both help (Dijkstra 1987) and harm

[6] Although misstate agreements would clearly be of concern, misstatement agreements occurred very rarely in both conditions; kappa for misstate agreements did not reach criterion largely because of the rareness of their observation.

(Weiss 1968) data quality. Current results point to the more frequent occurrence of digressions among both interviewers and respondents in Calendar interviews, and a greater level of laughter among interviewers in CQ interviews. The heightened laughter among CQ interviewers should not be taken as evidence that these interviews were more enjoyable to interviewers in comparison to the Calendar, as Belli et al. (2007) discovered that self-assessments from these same interviewers indicated greater enjoyment with the Calendar. Another possibility is to recognize that laughter may not be a sign of enjoyment as much as it is a sign of discomfort that may accompany a social interaction that is stilted due to the conventions of standardization (see Houtkoop-Steenstra 2000).

Finally, CQ interviews are noted to be less efficient than Calendar interviews in the more frequent asking of questions that add no new information (*nothing new*) to respondents' reports of their experiences. This loss of efficiency, however, should not be interpreted as translating into greater interviewing time for CQ interviews, as, in fact, Belli et al. (2007) found that the CQ interviews took 10% less time on average than Calendar ones.

4. Comparisons Across Studies

As expected, our current results with the CATI life course Calendar and CQ instruments replicate the pattern of findings that Belli et al. (2004; see Table 1 current study) observed with paper and pencil instruments testing a 2-year reference period, with some notable exceptions. Whereas the current study indicates that parallel or duration probing behaviors had been significantly more frequently used in the Calendar interviews, in Belli et al. (2004) duration probing occurred more frequently in CQ interviews. Moreover, whereas in Belli et al. (2004) timing probes occurred more frequently in the Calendar interviews, the current results reveal no significant differences in the use of timing probes between types of interviews, which is also true for sequential probes.

Differences in the occurrences of verbal behaviors between the current study and Belli et al. (2004) are primarily due to CQ design considerations. CQ designs are able to include scripted questions that seek sequential associations, although writing questions seeking parallel associations is prohibitive as it requires flexibility in order to identify candidate respondent experiences that have already been reported in the interview. In the CQ design studied by Belli et al., duration probes were frequently written in questions that asked respondents how long they had been employed, unemployed, or out of the labor force. In the CQ design that provides the current results, timing (e.g., "In what month and year did you move from < address > ?") and sequential probes (e.g., "What was the address of the place that you moved to?") were written into the scripted questions with skip patterns and "fills" in order to provide a means for respondents to provide comparable data with the Calendar design for purposes of testing between designs for data quality (see Belli, Smith, Andreski, and Agrawal 2007). The important point is that while Calendar designs will automatically encourage parallel and sequential probing, CQ designs are able to encourage certain types of sequential probing whenever such probing is built into the scripted questions.

Results with respondents are again in general correspondence with the results of Belli et al. (2004) in that parallel and sequential respondent retrieval strategies are more

prevalent in Calendar interviews. One notable difference is the overall frequency of the use of parallel strategies; whereas the use of respondent parallel strategies was infrequent – less than one occurrence per interview – in the Calendar instrument examined by Belli et al., there were frequent uses of spontaneous respondent parallel retrievals – an overall mean of twelve occurrences per interview – in the Calendar instrument that is the focus of the current results. In fact, the prevalence of parallel probes by Calendar interviewers was heightened in comparison to the interviewers who administered the Calendar examined by Belli et al., and respondents in the CQ interviews of the present study also had a notably higher use of parallel retrieval strategies compared to their counterparts who responded to the Belli et al. CQ instrument. The most likely reason for these differences resides in the length of the reference period. The two-year reference period in the instruments that were examined by Belli et al. did not lead to the remembering of many candidate events for purposes of parallel anchoring, whereas the life-course instruments that framed the current results apparently led to the recall of many candidate events. In addition, the frequent spontaneous use of parallel retrieval strategies in the life-course CQ instrument – an overall average of eight occurrences per interview – despite the infrequent parallel probing of interviewers, supports the notion that parallel associations are readily available in the structure of autobiographical memory.

Moreover, one other difference is that Belli et al. (2004) found no differences between Calendar and CQ interviews in rapport behaviors such as digressions and laughter exhibited by interviewers and respondents. However, current results illustrate more frequent occurrence of digressions among both interviewers and respondents in Calendar interviews, and a greater level of laughter among interviewers in CQ interviews. It is not clear why the results for rapport lack consistency with the results of Belli et al. One possibility is that reporting on the life-course provides more opportunities for interviewers and respondents to digress, and the more flexible nature of the Calendar interviews exacerbated the tendency to lose the focus on satisfactory questionnaire objectives.

5. Conclusion

In addition to promoting retrieval cues, Calendar interviewing has been consistently observed, because of a more flexible approach to interviewing in comparison to CQ methods, to encourage behaviors (1) in which conversation is used to ensure a shared meaning between interviewers and respondents and (2) that may exacerbate biased reporting among respondents. At this point in time it is unclear whether both retrieval cues and conversational benefits outweigh any increased risk of biasing respondents' answers as suggested by data quality studies that have revealed an overall advantage to Calendar instruments in comparison to CQ methods (see Belli and Callegaro 2009, for a review). One issue deserving consideration is that the greater flexibility that interviewers acquire in Calendar questionnaires appears to translate into a modestly increased impact that interviewers have on eliciting quality data from their respondents in comparison to CQs (Sayles, Belli, and Serrano 2010). Hence, the state-of-the-art of Calendar designs is one in which better overall data quality is compromised by some interviewers performing considerably better than others. It is likely that those interviewers who perform better in Calendar methods optimize both the use of retrieval probes and strategies and the

beneficial aspects of conversation while minimizing engaging in those verbal behaviors that potentially bias respondents' answers.

In conclusion, researchers and survey methodologists can rely on there being systematic differences between Calendar and CQ interviews in terms of the types of verbal behaviors that are engendered in each method. However, further research is necessary to investigate whether these systematic verbal behavior differences directly translate into a difference in data quality.

6. References

Balán, J., Browning, H.L., Jelin, E., and Litzler, L. (1969). A Computerized Approach to the Processing and Analysis of Life Histories Obtained in Sample Surveys. Behavioral Science, 14, 105–114.

Bartko, J.J. (1966). Intraclass Correlation Coefficient as a Measure of Reliability. Psychological Reports, 19, 3–11.

Beatty, P. (1995). Understanding the Standardized/Non-standardized Interviewing Controversy. Journal of Official Statistics, 11, 147–160.

Belli, R.F. (1998). The Structure of Autobiographical Memory and the Event History Calendar: Potential Improvements in the Quality of Retrospective Reports in Surveys. Memory, 6, 383–406.

Belli, R.F. and Callegaro, M. (2009). The Emergence of Calendar Interviewing: A Theoretical and Empirical Rationale. In Calendar and Time Diary Methods in Life-course Research, R.F. Belli, F.P. Stafford, and D.F. Alwin (eds). Thousands Oaks, CA: Sage.

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

Belli, R.F. and Lepkowski, J.M. (1996). Behavior of Survey Actors and the Accuracy of Response. Health Survey Research Methods: Conference Proceedings. 69–74. DHHS Publication No. PHS 96-1013.

Belli, R.F., Shay, W.L., and Stafford, F.P. (2001). Event History Calendars and Question List Surveys: A Direct Comparison of Interviewing Methods. Public Opinion Quarterly, 65, 45–74.

Belli, R.F., Smith, L., Andreski, P., and Agrawal, S. (2007). Methodological Comparisons Between CATI Event History Calendar and Conventional Questionnaire Instruments: Quality of Life Course Retrospective Reports. Public Opinion Quarterly, 71, 603–622.

Cannell, C., Miller, P., and Oksenberg, L. (1981). Research on Interviewing Techniques. In Sociological Methodology, S. Leinhardt (ed.). San Francisco: Jossey-Bass.

Cohen, J. (1960). A Coefficient of Agreement for Nominal Scale. Educational and Psychological Measurement, 20, 37–46.

Dijkstra, W. (1987). Interviewing Style and Respondent Behavior: An Experimental Study of the Survey Interview. Sociological Methods and Research, 16, 309–334.

Engel, L.S., Keifer, M.C., and Zahm, S.H. (2001). Comparison of Traditional Questionnaire with an Icon/Calendar-Based Questionnaire to Assess Occupational History. American Journal of Industrial Medicine, 40, 502–511.

Fleiss, J.L. (1971). Measuring Nominal Scale Agreement Among Many Raters. Psychological Bulletin, 76, 378–382.

Fowler, F.J. and Cannell, C.F. (1996). Using Behavior Coding to Identify Cognitive Problems with Survey Questions. In Answering Questions: Methodology for Determining Cognitive and Communicative Processes in Survey Research, N. Schwarz and S. Sudman (eds). San Francisco: Jossey-Bass.

Fowler, F.J. and Mangione, T.W. (1990). Standardized Survey Interviewing: Minimizing Interviewer-Related Error. Newbury Park, CA: Sage.

Freedman, D., Thornton, A., Camburn, D., Alwin, D., and Young-DeMarco, L. (1988). The Life History Calendar: A Technique for Collecting Retrospective Data. In Sociological Methodology, C.C. Clogg (ed.). San Francisco: Jossey-Bass.

Houtkoop-Steenstra, H. (2000). Interaction and the Standardized Survey Interview: The Living Questionnaire. Cambridge: Cambridge University Press.

Maynard, D.W. and Schaeffer, N.C. (2002). Standardization and Its Discontents. In Standardization and Tacit Knowledge: Interaction and Practice in the Survey Interview, D.W. Maynard, H. Houtkoop-Steenstra, N.C. Schaeffer, and J. van der Zouwen (eds). New York: Wiley.

Oksenberg, L., Cannell, C., and Kalton, G. (1991). New Strategies for Pretesting Survey Questions. Journal of Official Statistics, 7, 349–365.

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

Sayles, H., Belli, R.F., and Serrano, E. (2010). Interviewer Variance between Event History Calendar and Conventional Questionnaire Interviews. Public Opinion Quarterly.

Stafford, F.P. and Belli, R.F. (2009). Future Directions in Calendar and Time Diary Methods. In Calendar and Time Diary Methods in Life-course Research, R.F. Belli, F.P. Stafford, and D.F. Alwin (eds). Thousands Oaks, CA: Sage.

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, 232–241.

Weiss, C. (1968). Validity of Welfare Mothers' Interview Responses. Public Opinion Quarterly, 32, 622–633.

Yoshihama, M., Gillepsie, B., Hammock, A.C., Belli, R.F., and Tolman, R.M. (2005). Does the Life History Calendar Facilitate the Recall of Intimate Partner Violence? Comparison of Two Methods of Data Collection. Social Work Research, 29, 151–163.

Received April 2009
Revised November 2009