Journal of Official Statistics, Vol. 20, No. 2, 2004, pp. 185–218

Calendar and Question-List Survey Methods: Association Between Interviewer Behaviors and Data Quality

Robert F. Belli[1], *Eun Ha Lee*[2], *Frank P. Stafford*[3], *and Chia-Hung Chou*[4]

Insights are provided on the role of retrieval and conversational properties of Event History Calendars (EHCs) that promote higher quality retrospective reports than do standardized question-list (Q-list) methods. A verbal behavior coding analysis of 218 EHC and 197 Q-list interviews revealed significantly more behaviors in the EHC condition that indicated the use of timeline retrieval strategies and conversational engagement. Analyses of data quality measures demonstrate that there is not a significantly greater degree of interviewer variation on data quality in the EHC method. Better data quality was associated with a higher prevalence of retrieval cues, a greater degree of response openness, and lower levels of cognitive difficulty and rapport. The association of data quality and verbal behavior also interacted with method: retrieval cues and cognitive difficulty were directly associated with EHC response quality and indirectly associated with Q-list quality; rapport behaviors had a more detrimental effect on Q-list data quality.

Key words: Survey interviewing; response error; interviewer effects; interviewing style; autobiographical memory; retrospective reports; behavior coding.

1. Introduction

Event History Calendars (EHCs) have been shown to lead to higher quality survey retrospective reports than do traditional standardized question-list (Q-list) methods (Belli, Shay, and Stafford 2001; Yoshihama, Gillespie, Hammock, Belli, and Tolman 2003). Theoretically, EHCs are hypothesized to provide advantages with regard to data quality through a flexible conversational interviewing style that encourages respondents' narrative use of idiosyncratic retrieval cues available in the chronological and thematic structures of autobiographical memory (Belli 1998; Belli, Shay, and Stafford 2001). These cues are hypothesized to include top-down retrieval processes, in which specific details are cued with more abstract or general information, sequential retrieval processes, in which events or spells within the same autobiographical theme or domain are assessed as to their order of occurrence, and parallel retrieval processes, in which contemporaneous spells from more than one domain are used to provide greater precision in the timing of their respective occurrences (see also Barsalou 1988; Conway 1996).

[1] University of Nebraska, 238 Burnett Hall, Lincoln, NE 68588-0308, U.S.A. E-mail: bbelli@unl.edu
[2] University of Michigan.
[3] University of Michigan.
[4] University of Wisconsin.

Acknowledgment: This research has been supported in part by National Science Foundation grants SES-0001994 and SBR-9730297. We thank Wendy McCoy, Kristen Van Heest, and Ann Munster for their assistance during different phases of this project. We also thank three anonymous reviewers for their constructive comments.

Of these types of cues, Q-list interviews are typically designed to maximize top-down cues and provide only a restricted range of asking sequential retrieval cues. Because encouraging more complete reports of one's autobiographical past is optimized by the use of multiple cues that are available in interconnected associations among events, the advantage of EHC interviewing is hypothesized to reside in the introduction of parallel cues that are largely nonexistent in Q-list interviews, and the use of a greater variety of sequential cues (Belli 1998). Moreover, the flexible, more conversational style of EHC interviewing also may promote a greater ability to resolve uncertainties that are ubiquitously a part of verbal exchanges between conversants in their search for a shared meaning (Belli et al. 2001; Conrad and Schober 2000; Schober and Conrad 1997).

Supporting the beneficial role of cues in EHC interviews, Belli et al. (2001) found higher quality survey retrospective reports for an EHC condition that was tested in a direct experimental comparison against a traditional standardized Q-list method. The retrospective reports were obtained for a reference period of one to two years previously on variables that measure the quantity and frequency of economic behaviors. Specifically, the EHC outperformed the Q-list in eliciting reports on amount of income, and the number of weeks not working due to unemployment, the illness of oneself, and the illness of another. With these variables, the most consistent differences were obtained with the correlations between the EHC and Q-list reports and reports obtained in a standard of comparison, in which the EHC correlations were significantly stronger than those for the Q-list. As there was no difference between Q-list and EHC conditions in interviewing time, this improvement in data quality is in contrast to the findings of other studies to the effect that improvements from using timeline recall (Van der Vaart 2002) or conversational (Schober and Conrad 1997) aids involve increased costs associated with interviewing time.

Although the results of Belli et al. (2001) indirectly support the finding that the EHC increases the use of retrieval cues in comparison to Q-list interviews, and that these retrieval cues, in turn, promote higher quality retrospective reports, one aim of the present research is to directly examine whether retrieval cues are more prevalent in EHC interviews and to test for any beneficial effect of their use. An additional aim of the present research is to assess the role of other conversational processes, besides the use of retrieval cues, which can affect the quality of retrospective reports. In particular, we are interested in examining conversational behaviors that indicate that interviewers and respondents are 1) negotiating their uncertainty with regard to question and response meaning (Houtkoop-Steenstra 2000; Schaeffer, Maynard, and Cradock 1993) and 2) developing rapport in their interpersonal relationship (Belli, Lepkowski, and Kabeto 2001; Dijkstra 1987).

The strategy for assessing the role of retrieval cues in enhancing the quality of retrospective reports is to conduct a content analysis of verbal behaviors that indicate the presence of retrieval cues, and to associate specific instances of parallel, sequential, and top-down retrieval cues with data quality indices. The verbal behavior coding scheme that we applied to EHC and Q-list interviews was designed to be more comprehensive than extant research that has sought to discover associations of behavior with data quality in Q-list interviews (see, for example, Dykema, Lepkowski, and Blixt 1997; Belli and Lepkowski 1996). In verbal behavior coding designed to assess potential sources of survey error in Q-list interviewing, the focus is on determining whether interviewers are following the prescribed rules of standardized interviewing, including reading questions as written, probing nondirectively, and using appropriate feedback (e.g., Cannell and Oksenberg 1988). More recently, behavior coding has been assessing the quality of questions especially in the context of the extent to which respondents are expressing cognitive difficulties in answering questions as seen by their interruptions, qualified answers, expressions of uncertainty, and inadequate responses (Fowler 1992; Fowler and Cannell 1996; Oksenberg, Cannell, and Kalton 1991; Presser and Blair 1994). Although we are interested in examining all of these behaviors, largely because we need to account for types of behaviors that are more common in EHC than Q-list interviews, our comprehensive coding scheme also includes an assessment of behaviors associated with retrieval cues and other conversational processes that have not been included in earlier behavior coding schemes.

In addition to applying a comprehensive verbal behavior scheme to EHC and Q-list interviews to gain insight into both retrieval and conversational processes that may be responsible for improving the quality of retrospective reports in EHC interviews, a final aim of our research is to provide perspective regarding the overall utility of EHC interviewing. Because EHC methods advocate a flexible style of conversational interviewing, there is concern that the lack of standardization will compromise any improvements in data quality by an uncompromising cost of increased interviewer variation. By measuring the level of interviewer variation that is associated with data quality indices, we are able to provide an initial glimpse into whether EHC interviews show a marked disadvantage to Q-list interviews due to the flexible approach that interviewers assume in administering EHC designs.

2. Data Collection

All new methods and analyses reported in this article are derived from data collected in the experiment reported by Belli, Shay, and Stafford (2001). Because a detailed explanation of the data collection methods has already been provided in this earlier work, only a brief explanation is provided here. Readers who desire a more complete account of sample demographics, the exact question wordings and formats for both EHC and Q-list instruments, and the specific cues which interviewers were trained to use in order to enhance the quality of retrospective reports, are encouraged to consult Belli, Shay, and Stafford (2001).

Paper and pencil interviews were conducted via telephone during 1998 with a random subset of participants from the Panel Study of Income Dynamics (PSID) on events that occurred during the calendar years 1996 and 1997. PSID respondents constitute a nationally representative sample of United States households. Interviews were administered during a 6-week span from May through June of 1998. Respondents and 20 interviewers were randomly assigned to Event History Calendar (EHC; $N = 309$; 84.4% cooperation rate) and Question-list (Q-list; $N = 307$; 84.1% cooperation rate) conditions. The EHC instrument consisted of an 18×28-inch sheet that displayed timelines for the calendar years 1996 and 1997 on such topics as places of residence, members of the household, spells of being employed, unemployed, and out of the labor force, and receipt of ADC/AFDC and food stamps. A parallel Q-list instrument asked about the same topics within the same reference period, but used a traditional standardized interviewing format. With respondent permission, 95% of interviews were audio taped. Before their recruitment, all interviewers had received general interviewing training. In addition, interviewers in both conditions received 15 hours of study-specific training spread evenly over three days. For both conditions, training emphasized using interviewing techniques designed to maximize reporting accuracy that were appropriate for each condition. Interviewing minutes did not significantly differ between conditions (EHC $M = 17.1$, SE $= 0.55$; Q-list $M = 15.9$, SE $= 0.55$), $t(611) = 1.58$. Using data from the same respondents collected one year earlier during the 1997 PSID on reports of the same 1996 events as a standard of comparison, the quality of retrospective reports on 1996 events from the 1998 administration of EHC and Q-list interviews was assessed.

3. Verbal Behavior Coding

A verbal behavior-coding scheme was developed and applied to transcribed audio taped EHC and Q-list interviews. At the outset, we realized that there would be two main challenges in developing a reliable coding scheme. First, verbal behavior coding of EHC interviews had never been done before, and because there are no clearly defined question and answer sequences as in Q-list interviews, coders would need to make judgments concerning which utterances distinctly behave as specific types of questions/probes, feedback, clarification requests, and answers, in an undemarcated flow of verbal interchange. Second, as our aim was to develop a comprehensive coding scheme that could identify conversational and retrieval behaviors that could potentially affect data quality, we realized that the number of targeted behaviors would be quite large, and thus challenging to identify.

3.1. Coder training

Before the beginning of coder training, a coding team was organized that consisted of the first author, two professional interviewers, one who had experience with EHC interviewing (and who explicitly stated during training a preference for Q-list interviews), and two senior-level undergraduates, one majoring in psychology and the other in linguistics. Neither undergraduate had prior experience with interviewing or behavior coding. As preparation, the undergraduates read a number of articles on behavior coding and were exposed to behavior coding schemes developed for Q-list interviews.

Behavior coding training consisted of two phases. The first phase lasted ten weeks and concentrated on coding scheme development and the training of coders to use the coding scheme. At the start of training, an initial version of the coding scheme had been developed by the first author and introduced to the coders. This initial scheme included behaviors that had been identified in earlier research as being relevant to Q-list interviews (e.g., Fowler and Cannell 1996; Oksenberg, Cannell, and Kalton 1991), and the scheme also introduced retrieval and conversational behaviors that were hypothesized as being potentially important for both Q-list and EHC interviews (e.g., Belli, Lepkowski, and Kabeto 2001; Houtkoop-Steenstra 2000). In additional development of the coding scheme during the first phase, 2-3-hour weekly meetings were held that examined and evaluated 8 randomly selected transcripts, 5 EHC and 3 Q-list, for identifiable verbal behaviors.

The transcripts were coded as a group by the coding team, with new behaviors and the criteria for their coding being identified during the group meetings. Coding criteria were designed to provide some guidance regarding which sequences of words would consist of codable utterances, and there was considerable discussion on the parsing of exchanges into identifiable specific utterances that constituted particular behaviors. Contrary to the interaction coding scheme of Dijkstra and colleague (Dijkstra 2002; van der Zouwen and Dijkstra 2002), and consistent with the coding scheme developed by Cannell and colleagues (Cannell, Lawson, and Hausser 1975; Oksenberg, Cannell, and Kalton 1991), not all words were parsed into identifiable and codable utterances; coders were trained to restrict code assignments to only those utterances that qualified as identifiable behaviors according to the coding scheme (see Ongena 2002, for the distinction between full and selective coding). Some of the transcripts were reviewed at more than one meeting. Between meetings, coders reviewed the transcripts and the continuing progress of code development to assist in their learning of the emerging coding scheme. On average, each coder spent 7–8 hours per week engaged in coding activities (including the weekly meeting).

The second training phase lasted twelve weeks and focused on developing inter-coder reliability. In this phase, 12 randomly selected transcripts, 8 EHC and 4 Q-list, were independently coded and then brought to the group for discussion and for a qualitative assessment of inter-coder reliability. One transcript was independently coded and discussed each week. Weekly meetings again lasted from two to three hours. In these meetings, a few new behaviors were identified and the criteria for assigning codes were additionally refined. On average, each transcript required 4–5 hours for actual coding outside of the group meeting; thus coders were spending an average of eight hours per week during the second phase.

During this second session, it became apparent to the coding team that the professional interviewers, although having provided valuable input in coding scheme development, were having difficulty in assigning codes that were as reliable as those provided by the two senior-level undergraduates. Philosophically, one might expect that a valid coding scheme would require the coding of professional interviewers. However, our experience in this instance indicated that we were asking coders to recognize a high degree of behavioral detail in flexible and fluid speech that the professional interviewers found contrary to their preferred conception of interviews as being parsed into easily identifiable question and answer exchanges, consistent with their experience of administering standardized Q-list interviews. Accordingly, only the undergraduates continued to serve as coders following the termination of the weekly meetings.

After all group meetings and before production coding, each undergraduate independently coded a new set of 4 randomly selected transcripts, 2 EHC and 2 Q-list, which were then qualitatively evaluated by the first author so as to determine whether there existed sufficient inter-coder reliability for production coding to commence. The coders received no feedback from the first author for this activity. As there were only a few transcripts being evaluated, reliability measures for individual codes were not assessed at this time. Because actual transcripts were examined in this evaluation, the first author was able to examine code notations for each speaking turn by the interviewers and respondents. The agreement in code assignments between coders for each turn appeared strong, although not perfect. Similar to the observations of Cannell, Lawson, and Hausser (1975, p. 29) in their coding of Q-list interviews, disagreements appeared to be mostly the result of one coder observing a behavior that the other coder did not, rather than the same or similar utterances being identified as two different behaviors. The overall assessment by the first author was that the undergraduates were adequately consistent in their code assignments to commence with production coding.

3.2. Production coding and reliability analysis

Attesting to the comprehensiveness of the coding plan, 56 verbal behaviors were targeted for behavior coding in the final coding scheme. During training and production coding, both EHC and Q-list interviews were segmented into seven domains, which corresponded to questions on 1) residence, 2) household composition, 3) employment, 4) earned income, 5) unemployment and out of the labor force, 6) time away from work, and 7) entitlements. Another domain, landmarks, was unique to the EHC. Due to skip patterns in both EHC and Q-list interviews, not all respondents were asked about every domain.

Coders treated exchanges within a domain as a unit of analysis, with codes assigned sequentially in the order of the behaviors that were observed to occur within each domain. Any particular turn taken within an interviewer and respondent exchange was composed of one or more utterances. With the exception of the assignment of *directive*, any uniquely identified utterance was assigned only one behavior. A few behaviors, such as *how long ago* and *refused answering*, were not observed during training sessions but were retained as part of the coding scheme because of their potential importance.

Because of refusals, inaudible tapes, and attrition of the coders to other career opportunities (both were accepted for graduate school) before all targeted interviews were coded, 218 of the 309 EHC interviews, and 197 of the 307 Q-list interviews, were coded. Of these 415 interviews, 38 (9.2%) interviews (17 EHC and 21 Q-list) were independently coded by both coders in order to measure the reliability of verbal behavior code assignments. These 38 double-coded transcripts were not previously examined in any weekly meetings, but they did include the 4 transcripts that had been previously double-coded right before production. Both coders were unaware of which transcripts were being double-coded during the production phase. There was no intermittent evaluation or analysis of double-coded transcripts during the production phase, and coders received no feedback.

In the reliability analyses of the 38 double-coded interviews, Pearson correlations between coders for each behavior were calculated on the basis of the frequency of code assignments within each domain and within each interview. Specifically, each domain was treated as a unit of analysis, with the counts of each code assigned to each domain being tallied for each coder, and a Pearson correlation between coders calculated on the frequencies of these code assignments, across each of the 38 transcripts. Thus, for each calculated correlation there were 265 data points, which were derived from the 17 EHC interviews and their mean of 7.59 domains, and the 21 Q-list interviews and their mean of 6.48 domains $((17 \times 7.59) + (21 \times 6.48)) = 265$. Because of the exploratory nature of this research, a liberal criterion of an $r \geqq .40$, demonstrating at least a moderately strong association, was used to identify those codes with sufficient inter-coder reliability to warrant additional analyses. Forty of the 56 behaviors met this criterion.

Typically, reliability analyses in Q-list behavior coding are based on kappa statistics that use the question-answer exchange as the unit of analysis (see Oksenberg, Cannell, and Kalton 1991), a unit that is considerably more fine-grained than using the domain as the unit of reliability analysis. In Q-list interviews, the kappa statistic–based on *whether or not* coders agree on at least one of the same codes being assigned to each of the question-answer exchanges–is appropriate as there are likely only to be zero, one, or at most a few observations of any given behavior within any question-answer exchange. Of course, because EHC interviews do not have predefined question-answer sequences, a finer-grained unit other than the domain is not readily available. As a result, at times coders do observe many of the same behaviors within a given domain, as the amount of interviewer-respondent interaction within a domain can be quite large. We selected the Pearson correlation as our measure of inter-coder reliability as it is sensitive to differences in magnitude, and because it is also appropriate for interval-level continuous data to which the upward bound is unknown.

An important potential problem of using the domain as the unit of analysis is that spuriously high correlations could result even when there is a lack of agreement between coders as to which particular utterances constitute which specific behaviors, just as long as the number of specific code assignments is consistently close in number. We have no direct evidence to eliminate this concern from contention. However, judging by observations that occurred during coder training and the qualitative evaluation of the 4 post-training double-coded transcripts, the majority of inter-coder agreements are likely the result of both coders treating similar sequences of words as the same behavior.

Table 1 presents the verbal behavior coding scheme and definitions for the 40 behaviors that fulfilled the criterion. A number of different interviewer verbal behaviors were identified, corresponding to a) retrieval, b) calendar year questions/probes, c) uncertainty behaviors, d) problems with standardization, e) a response to perceived respondent cognitive difficulty, f) feedback, and g) rapport. Respondent behaviors were identified as including h) retrieval strategies, i) a response to interviewer uncertainty, j) cognitive difficulty, k) miscellaneous behaviors, l) rapport, and m) a response to interviewer narrowing probe.

Although not identified reliably, the *parallel probe* behavior was considered to be of such substantive importance that an additional pass was conducted to determine the reason for coder discrepancies. Discussions between the first author and the coders resulted in the determination that one coder had missed recognizing many parallel probes. Audiotapes were reexamined exclusively for parallel probe behaviors, and additional utterances were identified. In addition, parallel probes in the double-coded transcripts were recoded to be completely consistent between the two coders. Column 2 of Table 2 provides the inter-coder reliabilities for the 41 behaviors that were subjected to additional analyses. As can be seen, reliabilities for most codes were substantively stronger than the minimum criterion of $r \geq .40$.

Table 1. Verbal behavior codes and descriptions

Code	Description
	A. Interviewer retrieval questions/probes
	A.1. Parallel retrieval
Holiday probe	Interviewer uses a public holiday as an anchor for probing on the timing of a spell.
Parallel probe	Interviewer uses an event from the respondent's past as an anchor for probing on the timing of a spell in a different domain. Example: *How about the fall and the winter, when your brother and sister came to live with you. Were you ever laid off when they were living with you?*
	A.2. Sequential retrieval
Duration	Interviewer is seeking how long, how much time, or in which months, a spell occurred. Example: *How long were you at that address?*
Continuity	Interviewer is seeking clarification regarding whether a spell continued during a specified period. Example: *Have you moved any time since the spring of 1996?*
Timing	Interviewer is seeking information on the beginning or ending of a spell. Example: *When did she graduate?*
Time gap fill	Gap exists in timing of spells that interviewer is trying to resolve. Example: *Did you have any time off in between, or you just went straight to (company name)?*
	A.3 Top-down retrieval
Top-down	Interviewer is seeking data elements for a spell that is already identified as occurring within a calendar year. Code only once for any continuous series of top-down exchanges. Example: *What was your address there?*
	A.4 Miscellaneous retrieval
Balanced narrowing	Interviewer is refining transition point by inquiring with all possible temporal reference points within a broader temporal category. For example, when narrowing from month to thirds of the months, all the thirds-of-month are included in the query, such as: *Did you move in the beginning, at the end, or sometime in the middle of April?*
Unbalanced narrowing	Interviewer is refining transition point by inquiring with a restricted number of all possible temporal reference points within a broader temporal category. For example, when narrowing from month to thirds of the months, less than all three thirds-of-month are included in the query, such as: *Did you move in the beginning of April?*
How many	Interviewer asks about the frequency of entities. Example: *How many jobs do you currently have?*
Ever	Interviewer is seeking information on new spell without defining a reference period. Example: *Have you ever done work for pay?*

Table 1. Continued

Code	Description
Example	An example of the type of possible spell or transition is offered. Example: *Perhaps, you know, like I said, maybe a special vacation or a family reunion?*
Preload	Interviewer is using preloaded information. Preloaded information can include last known address and household members as of the last interview.
	B. Interviewer calendar year questions/probes
Free year	Interviewer is seeking information on new spell allowing respondent to choose to start at any year within a defined reference period. Example: *Are there any events in the past few years, from the end of 1995 to the present that stand out in your mind?*
Forced year	Interviewer is seeking information on new spell by forcing respondent to begin at a particular year. Example: *Did you take any vacation or time off during 1996?*
	C. Interviewer uncertainty behaviors
Interviewer verification	Interviewer verifies with the respondent information that has already been provided. This is not a repetition of a response, as there is a clear indication that interviewer has some uncertainty about the response.
Interviewer seeks clarification	Interviewer seeks clarification from respondent on some aspect of the survey/questionnaire in which information beyond that directly provided by the respondent is sought.
	D. Interviewer problems with standardization
Significant change	A significant change in question wording is one that appears to change, or can conceivably change, the meaning of a question. For Q-list interviews only.
Wrong skip	Interviewer asks a wrong question or one that does not apply. For example, in EHC, the respondent was never employed but is still asked time away questions (sick, vacation, etc.) In Q-list, a wrong skip pattern was followed by the interviewer.
Directive	Any probe can also be coded as directive (except narrowing probes and interviewer verification), if it poses the risk of biasing the respondent's answer.
	E. Interviewer response to perceived cognitive problem
Interviewer clarification	Interviewer provides clarification on some aspect of the survey/questionnaire.

Table 1. Continued

Code	Description
	F. Interviewer feedback
Acceptable feedback	Neutral phrases following question answering that are either short or long showing appreciation for receipt of the response. Example: *Thank-you.*
Unacceptable feedback	Nonneutral phrases following question answering that are either short or long showing appreciation for receipt of the response. Example: *Oh shoot. It's been a tough year for you.*
Task related feedback	Interviewer refers to some logistical task-related component of the interviewing process. Example: *Ok, if you just wait a second, I'll just mark this down here.*
	G. Interviewer rapport behaviors
Interviewer digression	Interviewer asks a question or makes a comment that is not a direct attempt to satisfy study or question objectives. Example: *So you're satisfied with your job?*
Interviewer laughs	Interviewer laughs.
Scripted distancing	Within the scripts given to interviewers to read, interviewer makes a comment that provides information to the respondent that the questions originate with a third party, the survey researcher. Example: *My instructions are to ask these questions of everybody.*
	H. Respondent retrieval strategies H.1. Parallel retrieval
Spontaneous parallel	Respondent spontaneously refers to a contemporaneous event in a domain different than the one inquired by the interviewer. Example: *It was football season when it started up.*
	H.2. Sequential retrieval
Spontaneous sequential	Respondent spontaneously provides the duration, continuity, or timing of a spell that is not directly probed by interviewer. Example: *Now we're divorced.*

Table 1. Continued

Code	Description
	I. Respondent response to interviewer uncertainty
Agreement	Respondent agrees with interviewer verification.
	J. Respondent cognitive difficulty behaviors
Request for clarification	The respondent indicates that more information is needed to answer the question, including requests that the question be repeated.
Does not meet	Response is an attempt to answer the question but fails to meet question objectives.
Correction	Respondent corrects a response to a previous question.
Don't know	Don't know response.
	K. Miscellaneous respondent behaviors
Nothing new	Response indicates that there is no new spell nor new top-down data to enter into survey instrument.
Same information	Respondent says that the characteristics reported for one entire calendar year are the same as for the other entire calendar year.
Third party	A person other than the respondent at the respondent's household assists with answers.
	L. Respondent rapport behaviors
Respondent digression	Respondent makes a comment that is not a direct attempt to satisfy study or question objectives. Example: *I don't remember things like I did before*.
Respondent laughs	Respondent laughs.
	M. Respondent response to unbalanced narrowing probe
Option selected	Respondent selects an option interviewer provides.
Option not selected	Respondent does not select an option interviewer provides.

Table 2. Intercoder reliability (N = 38) and mean code assignments for EHC (N = 218) and Q-list (N = 197) conditions

1. Behavior	2. Intercoder reliability (*r*)	3. Condition mean (SD)		4. *t*-test
		EHC	Q-list	
Interviewer behaviors				
Holiday probe	.88	0.48 (1.19)	0.00 (0.00)	5.91*
Parallel probe	1.00	0.51 (1.04)	0.06 (0.29)	6.22*
Duration	.91	2.18 (2.17)	4.25 (3.25)	−7.54**
Continuity	.88	3.73 (2.88)	3.19 (2.03)	2.25
Timing	.88	6.62 (4.61)	4.76 (3.08)	4.88*
Time gap fill	.71	0.16 (0.44)	0.02 (0.12)	4.71**
Top-down probe	.93	9.89 (5.97)	11.67 (7.44)	−2.67
Balanced narrowing	.98	1.60 (1.69)	0.015 (0.12)	13.81**
Unbalanced narrowing	.77	0.68 (1.15)	0.02 (0.17)	8.37**
How many	.95	0.03 (0.20)	1.11 (0.93)	−15.83**
Ever	1.00	0.01 (0.10)	0.33 (0.51)	−8.64**
Example	.89	1.70 (1.41)	0.01 (0.07)	17.72**
Preload	.85	1.84 (1.75)	1.97 (1.57)	−0.74
Free year	.89	15.19 (4.72)	1.24 (0.69)	43.12**
Forced year	.94	11.53 (7.56)	29.53 (8.97)	−21.98**
Interviewer verification	.75	11.30 (11.21)	3.68 (5.45)	8.93**
Interviewer seeks clarification	.85	13.22 (11.41)	8.19 (7.97)	5.25**
Significant change	.64	0.01 (0.15)	4.56 (4.62)	−13.80**
Wrong skip	.47	0.25 (0.50)	0.79 (1.09)	−6.40**
Directive	.49	2.10 (2.73)	1.43 (2.49)	2.60
Interviewer clarifies	.83	2.44 (2.38)	2.88 (3.32)	−1.51
Acceptable feedback	.96	8.22 (7.22)	6.12 (9.09)	2.58
Unacceptable feedback	.71	3.58 (3.91)	1.10 (1.35)	8.80**
Task related feedback	.63	2.11 (2.35)	0.83 (1.47)	6.74**
Interviewer digression	.66	2.38 (2.49)	2.26 (2.87)	0.42
Interviewer laughs	.96	2.17 (2.90)	1.82 (3.36)	1.12
Scripted distancing	.97	1.17 (1.34)	5.11 (2.39)	−20.49**

Table 2. Continued

1. Behavior	2. Intercoder reliability (*r*)	3. Condition mean (SD)		4. *t*-test
		EHC	Q-list	
Respondent behaviors				
Spontaneous parallel	.42	0.17 (0.51)	0.05 (0.23)	3.34*
Spontaneous sequential	.60	1.81 (2.13)	0.68 (1.20)	6.72**
Agreement	.76	10.67 (10.38)	3.50 (5.16)	9.03**
Request for clarification	.90	3.27 (3.04)	3.38 (3.62)	−0.32
Does not meet	.46	2.14 (2.19)	1.88 (2.51)	1.12
Correction	.73	0.91 (1.27)	0.70 (1.10)	1.87
Don't know	.66	1.04 (1.71)	0.67 (1.02)	2.71
Nothing new	.91	17.06 (5.02)	21.14 (5.66)	−7.78**
Same information	.57	0.26 (0.56)	0.27 (0.62)	−0.30
Third party	.77	0.19 (0.77)	0.12 (0.53)	1.03
Respondent digression	.74	4.88 (4.99)	3.66 (4.53)	2.60
Respondent laughs	.94	1.35 (2.70)	1.30 (2.37)	0.21
Option selected	.66	0.39 (0.73)	0.01 (0.10)	7.68**
Options not selected	.86	0.19 (0.50)	0.01 (0.07)	5.50**

*$p < .01$; **$p < .001$; adjusted using Holm's sequentially rejective multiple Bonferroni test procedure.

3.3. Differences in behaviors between EHC and Q-list conditions

Summing observations across domains, Column 3 of Table 2 reveals the mean number and standard deviation of each behavior per interview separately for EHC and Q-list conditions. In addition, Column 4 of Table 2 depicts the results of *t*-tests designed to examine whether specific behaviors significantly differed in the frequency of their occurrence between conditions. Whereas certain behaviors occurred frequently, others occurred rarely. In addition, of the 41 behaviors, 25 prove to have significantly differed in their frequency between conditions, at $\alpha = .05$, using Holm's (1979) sequentially rejective Bonferroni test procedure to control for Type I errors.

3.3.1. Retrieval behaviors

Of specific interest were those behaviors hypothesized to benefit the retrieval process. Belli (1998) highlights parallel, sequential, and top-down retrieval processes as those that could encourage greater precision in retrospective reports. Of these processes, EHC interviews are expected to encourage the use of more extensive parallel and sequential retrieval processes in comparison to Q-list ones. Interviewer probing that signifies the encouragement of parallel retrieval processes is represented by the *holiday probe* and *parallel probe* behaviors. Both of these behaviors signify the use of parallel probing on the part of interviewers as they both represent the instantiation of a contemporaneous spell in another domain to aid in the remembrance of a to-be-remembered spell. In addition, the *spontaneous parallel* behavior represents utterances in which respondents spontaneously made parallel retrieval attempts. As for sequential retrieval processes, *duration*, *continuity*, *timing*, and *time gap fill* behaviors are ones in which respondents would either directly or indirectly have information relevant to determining whether spells occurred before or after other spells within the same domain. In addition, the *spontaneous sequential* behavior provides an indication of situations in which respondents spontaneously engage in sequential retrieval attempts. Finally, *top-down* retrieval indicates situations in which interviewers were probing respondents to provide detailed information for spells of behavior that had already been defined during the interview, such as asking for average hours worked per week for a particular spell of employment.

With the exception of *time gap fill*, an examination of the mean number of behaviors reveals that the sequential retrieval behaviors occurred much more frequently than those behaviors indicative of parallel retrieval. As interviewers in the EHC condition were explicitly trained in the use of parallel retrieval probes, the rare observation of *parallel* behaviors, which only occur on average in one of two interviews, indicates that the interviewer training was not as successful as hoped. Yet, in comparing the EHC and Q-list conditions for all of the sequential and parallel interviewer probes and spontaneous respondent retrievals, with the exception of *duration* and *continuity*, each appeared significantly more often in the EHC than in the Q-list interviews. As for *duration* behaviors, they appeared more often in Q-list than EHC interviews. *Top-down* probes did not differ in their frequency of use between conditions. As expected, interviewers in the EHC condition were more frequently probing with the use of a variety of parallel and sequential retrieval techniques that were expected to improve the quality of reports, and respondents were also more likely to spontaneously use these techniques in the EHC condition.

Additional retrieval behaviors that deserve mention are *balanced narrowing*, *unbalanced narrowing*, and *how many*. Not surprisingly, *balanced narrowing* and *unbalanced narrowing* are more prevalent in EHC than Q-list interviews as they represent techniques emphasized in EHC training on how to probe for more precise timing of spell transitions after the respondent has identified a less refined estimate. *How many* is especially noteworthy as encouraging respondents to engage in a behavioral frequency report, and its occurrence is more frequent in Q-list than in EHC interviews. As EHC interviewing is designed to seek information within a reference period on when events happened, instead of how many events happened, the very low prevalence of *how many* behaviors in the EHC interviewers is not surprising.

3.3.2. Calendar year probes

The *free year* and *forced year* behaviors distinguish between spells that were queried by interviewers to either permit the respondent to choose which calendar year in the 1996 to 1997 reference period to start with (*free year*), or to permit the respondent to start with a particular calendar year (*forced year*). By design, the Q-list interviews required interviewers to specify a single calendar year, either 1996 or 1997 for most domains (randomly assigned), to facilitate the collection of calendar year estimates of number of jobs, weeks per year, income, and entitlement receipts. As for the EHC, during training interviewers were encouraged to allow respondents to choose which calendar year they preferred to begin their reports, especially for the landmark, employment, unemployment and out of the labor force, time away, and entitlements domains. Because calendar year estimates from EHC interviews could be constructed after data collection by decomposing the two-year timeline, EHC interviews permitted more latitude with regard to the calendar year starting point. In addition, as the psychological literature suggests that allowing individuals to choose the chronological direction of retrieval leads to more accurate remembering than forcing a forward or backward chronological retrieval (Loftus and Fathi 1985; Jobe, White, Kelley, Mingay, Sanchez, and Loftus 1990), study staff determined that a free order was advisable in EHC interviews whenever possible, and EHC interviewers were trained accordingly. As expected, the EHC condition led to more frequent *free year* behaviors than the Q-list, whereas the Q-list interviewers provided significantly more *forced year* behaviors than EHC interviewers.

3.3.3. Interviewer uncertainty behaviors

Fairly often, interviewers would express some degree of uncertainty concerning the information presented to them, as evidenced by the *interviewer verification* and *interviewer seeks clarification* behaviors. Both occur significantly more often in EHC than Q-list interviews. In addition, the *agreement* behavior signified whether respondents agreed with an *interviewer verification*, and as a direct response to *interviewer verification*, is also observed to occur more frequently in the EHC interviews.

3.3.4. Problem behaviors

Several authors (Belli, Lepkowski, and Kabeto 2001; Fowler 1992; Fowler and Cannell 1996; Oksenberg, Cannell, and Kalton 1991) have identified a number of verbal behaviors as problematic because they appear to threaten the quality of responses. Some behaviors are problematic because they violate maxims of standardization, others are problematic because they are signs that respondents are experiencing cognitive difficulty. Regarding the former type of behavior, *significant change to question wording* was to be assigned only to Q-list interviews (as interviewers were free to paraphrase introductory scripts in the EHC), *wrong skip*, although infrequent, occurred more frequently in Q-list than EHC interviews, and the frequency of *directive* behaviors did not significantly differ between EHC and Q-list interviews. Regarding behaviors that indicate respondent cognitive difficulty, *request for clarification*, *does not meet question objectives*, *correction*, and *don't know* behaviors did not differ in their frequency between conditions. Also of note is the observation of *interviewer clarifies* and *nothing new* behaviors. Interviewers likely provide clarification when they perceive respondents to be experiencing cognitive difficulty, and *interviewer clarifies* behaviors occur equally often in EHC and Q-list interviews. *Nothing new* was assigned to indicate the occurrence of inquiries that did not add any new information about spells or data elements. Although frequently observed in both conditions, *nothing new* is observed to occur more often in Q-list than EHC interviews, signifying a lesser degree of efficiency in Q-list questioning than in the EHC. Such inefficiency, and other inefficiencies in the Q-list such as redundancy in respondents being asked to provide the same employer name in more than one calendar year section of the questionnaire (when applicable), no doubt contributed to increasing interviewing time in the Q-list in comparison to the EHC.

3.3.5. Interviewer feedback and rapport behaviors

In survey interviews, interviewers often provide feedback to respondents concerning the perceived adequacy of the response process, and both participants engage in behaviors whose content is outside the task of satisfying survey objectives, often as a means of developing rapport. We included the observation of *acceptable feedback* and *unacceptable feedback* behaviors; *unacceptable feedback* was observed to occur significantly more frequently in EHC than in Q-list interviews. In addition we included the observation of *task related feedback:* the more frequent occurrence of this behavior in the EHC interviews suggests that interviewers were taking more time, or expending more effort, to record responses in the EHC than the Q-list, and felt that they needed to fill in potential periods of silence. As for rapport behaviors, *interviewer* and *respondent digressions* and *laughter* occur fairly frequently in both conditions, and neither occurs significantly more frequently in one condition than the other. We also observed *scripted distancing*, in which interviewers, by design, were explicitly indicating that the questions originated from a third party, the survey researcher. *Scripted distancing* behaviors are observed to occur more frequently in Q-list than in EHC interviews.

3.4. Excluded behavior codes

Table 3 lists the 16 targeted behaviors that were excluded from additional analyses, their inter-coder reliability statistics, and the number of unique identifications across the 38 double-coded interviews. Almost all of these behaviors were assigned codes infrequently, and the lack of exposure of the coders to these behaviors no doubt contributed to a lack of reliability. Three behaviors have only one unique identification in the double-coded transcripts, causing the computation of correlation to be undefined. When it comes to behaviors with at least 17 unique coding assignments, all but one of these yield a positive correlation. The lone exception is a code designed to identify verbal behaviors in which interviewers were *distancing* themselves from authorship of the query being asked (see Houtkoop-Steenstra, p. 52), and the lack of reliability for this identification is largely due to the undergraduate majoring in linguistics having identified many more of these behaviors than the other undergraduate coder. Although coded reliably, two behaviors, *refusals* and *response made*, are excluded from analyses for reasons that are particular to each behavior. There are only two unique identifications (and where the two coders agreed) for *refusals*, and because of the low frequency of identification, the reliability of the coding is less than certain. *Response made* is excluded from analyses because it merely served as a default code for any utterance by respondents to which other codes did not apply. In other words, *response made* simply served the purpose of aiding coders to maintain an awareness of utterance distinctiveness and structure, and the flow of information from respondent to interviewer, and thus to prime coders' sensitivity to demarcating other ensuing respondent or interviewer behaviors. In sum, the excluded behaviors do not reflect any profound tendency for the coders to have assigned codes unreliably.

Table 3. Intercoder reliabilities and number of unique identifications for the 38 double-coded transcripts for behaviors excluded from analyses

Behavior	Intercoder reliability (*r*)	Unique identifications
Interviewer behaviors/		
How long ago	Undefined	1
Parallel probe	−.01	6
Study concept	.33	17
Distancing	.03	42
Respondent behaviors/		
Refused answering	1.00	2
Spontaneous holiday	−.00	2
Spontaneous top-down	.21	18
Spontaneous how long ago	Undefined	1
Spontaneous sequential	Undefined	1
Rate	−.00	3
Count or enumerate	Undefined	5 (all by one coder)
External reference	Undefined	4 (all by one coder)
Qualified answer	.34	48
Irritation	−.00	4
Disagreement	−.02	10
Response made	.97	1,385

3.5. Overall impressions

There is no doubt that EHC and Q-list conditions encouraged different interviewing styles. The EHC condition is marked by more frequent use of a variety of retrieval strategies that because of their affiliation to the structure of autobiographical memory are hypothesized to benefit the quality of retrospective reports. In addition, differences in the frequencies of behaviors between conditions reflect a more flexible interviewing style in the EHC condition, with the Q-list being more constrained in style as a standardized interviewing approach. Participants in the EHC interviews are observed to be more conversationally engaged than those in the Q-list interviews, as seen by the larger frequency of interviewers expressing uncertainty and seeking clarification from respondents, and by a greater degree of interviewer unacceptable feedback.

4. Interviewer Variation

Because one concern of the flexible style of conversational interviewing that typifies EHC methodologies is greater dependence, in comparison to standardized Q-list interviewing, on the skill of interviewers to promote quality respondent reports, we conducted analyses to determine whether the EHC and Q-list interviewing methods result in significantly different levels of interviewer variation in response quality. Kish (1962) introduced the concept of the intraclass correlation, ρ, to denote error variation associated solely with interviewers; ρ is also referred to as a measure of interviewer effects (Groves 1989, p. 318). Following the Kish model, we calculated ρ on each of the signed and absolute difference measures of data quality appearing in Belli, Shay, and Stafford (2001). These data quality measures are based on reports of income, weeks working, weeks out-of-the labor force, weeks unemployed, and weeks missing work due to vacation, the illness of oneself, and the illness of another. There are also two composite variables: total illness combines reports of weeks missing work from self-illness and other-illness, and total weeks away includes both of these illness measures as well as weeks on vacation. In cases in which ρ was negative, following procedures adopted by other researchers (Bailar, Bailey, and Stevens 1977; Mangione, Fowler, and Lewis 1992), we recoded these values to a small positive value (.001). The values of the resulting ρ's for each data quality measure range from .001 to .030 in the EHC condition, and from .001 to .047 in the Q-list condition.

One weakness with these computations of ρ is the violation of the assumption of the random assignment of respondents to interviewers. Although interviewers and respondents were randomly assigned to EHC and Q-list conditions, the assignment of interviewers to respondents was affected by production demands. Another weakness centers on a lack of power due to the small number of interviewers, and the different number of interviews conducted by each interviewer in the Q-list ($n = 10$ interviewers who conducted a mean of 28.0 interviews with a range between 10 and 42 interviews) and EHC ($n = 9$ interviewers, who conducted a mean of 31.8 interviews with a range between 13 and 53 interviews) conditions.

To determine whether the EHC and Q-list conditions led to different levels of interviewer effects, we conducted significance tests between conditions separately for the signed and absolute difference measures of ρ. The mean levels of ρ in the EHC condition are nonsignificantly lower than those in the Q-list condition for both signed

($t(16) = -0.82$, $p = .43$) and absolute differences ($t(16) = -1.73$, $p = .10$). As the normality assumption is likely violated by the skewness of the ρ distributions, we conducted Wilcoxon signed ranks tests, which also resulted in no differences between the conditions: $z = -0.52$, $p = .60$ for signed differences, and $z = -0.56$, $p = .58$ for absolute differences.

As Belli, Shay, and Stafford (2001) also conducted analyses based on the correlations between experimental (1998 EHC and Q-list) and standard of comparison (1997 PSID) reports, we sought to determine whether there exist condition differences in interviewer variability based on correlations as the outcome measure. However, a measure of interviewer variability in this context is less than straightforward. Kish's intraclass correlation, ρ, is based on using data from each interview as the unit of analysis; with correlation coefficients, each interviewer's set of interviews is the smallest possible unit of analysis. Accordingly, we computed separately for each interviewer the correlation that they had obtained between their experimental reports and the reports in the standard of comparison for each outcome variable. Next, for each outcome variable, we determined whether the variance among interviewers in the obtained correlations for each outcome measure differed between conditions as measured by a Levene test. Only for unemployment is there a significant difference in variation between conditions, and the larger variation in interviewer correlations occurs in the Q-list condition.

5. Behaviors and Data Quality

Finding no evidence of larger interviewer effects in EHC interviews than in Q-list ones, we next conducted analyses focused on determining those behaviors, if any, which show associations with data quality. To reduce the number of behaviors that require analysis to a manageable level, and to account for the clustering of certain behaviors as consistently appearing together within interviews, we conducted a principal components analysis of observed behaviors per interview. Based on the factors obtained in the principal components analysis, we then conducted analyses to determine whether associations exist between verbal behavior factors and measures of data quality, and whether any associations are dependent on EHC or Q-list interviewing methods.

5.1. Principal components analysis

A principal components analysis following a varimax orthogonal rotation results in four latent factors with eigenvalues larger than 2; these four factors also demonstrate conceptual coherence by representing interpretable latent variables. Table 4 reveals the factor loadings of the behaviors. Each factor was identified by those behaviors with factor loadings ≥.40, and when a behavior loaded at .40 or higher on more than one factor, with one exception the behavior was assigned to the factor to which the loading was highest. The one exception involved *spontaneous sequential*, which was assigned to one factor with a factor loading of .40 despite having a factor loading of .41 with another factor. As these loadings were nearly identical in value, *spontaneous sequential* was assigned to the factor that conceptually consisted of behaviors that had more characteristics in common.

The four factors are named **Retrieval Cues**, **Detailed Interviewing**, **Cognitive Difficulty**, and **Rapport**. The **Retrieval Cues** factor is populated by retrieval behaviors,

Table 4. Rotated orthogonal varimax factor pattern

Behavior	Retrieval cues	Detailed interviewing	Cognitive difficulty	Rapport
Interviewer behaviors				
Holiday probe	***58***	−11	−04	−03
Parallel probe	***56***	−11	12	07
Duration	20	***63***	47	−05
Continuity	***50***	11	15	−04
Timing	***61***	10	45	15
Time gap fill	33	−14	14	13
Top-down probe	40	***53***	51	11
Balanced narrowing	***57***	−37	16	−05
Unbalanced narrowing	21	−15	15	−01
How many	−15	***79***	22	−02
Ever	−28	05	−39	01
Example	41	***−47***	04	−07
Preload	33	27	−08	−07
Free year	43	***−70***	19	03
Forced year	02	***92***	13	01
Interviewer verification	***81***	−06	12	28
Interviewer seeks clarification	16	−14	***64***	36
Significant change	−21	***61***	02	29
Wrong skip	−04	***48***	−11	11
Directive	***49***	20	16	13
Interviewer clarifies	−03	19	***68***	26
Acceptable feedback	04	−15	***52***	−05
Unacceptable feedback	21	−25	13	***49***
Task related feedback	***56***	−03	03	07
Interviewer digression	04	08	12	***76***
Interviewer laughs	05	06	02	***71***
Scripted distancing	−06	***84***	−16	−15

Table 4. Continued

Behavior	Retrieval cues	Detailed interviewing	Cognitive difficulty	Rapport
Respondent behaviors				
Spontaneous parallel	***52***	02	− 13	02
Spontaneous sequential	***40***	− 17	11	41
Agreement	***81***	− 06	13	28
Request for clarification	09	19	***68***	24
Does not meet	10	06	***62***	17
Correction	34	07	26	32
Don’t know	23	01	29	− 01
Nothing new	23	***65***	13	− 11
Same information	− 10	− 01	35	− 12
Third party	03	− 04	11	00
Respondent digression	05	− 09	18	***83***
Respondent laughs	12	11	03	***61***
Option selected	21	− 15	10	− 01
Options not selected	10	− 11	04	− 03
Eigenvalue	7.47	5.67	2.82	2.05

including those associated with parallel retrieval (*holiday probe, parallel probe, spontaneous parallel*), sequential retrieval (*continuity, timing, spontaneous sequential*), and the *balanced narrowing* probe. *Interviewer verification* also tends to occur alongside behaviors that are indicative of retrieval attempts; apparently interviewers seek to verify their respondents' retrieval attempts and the success of the strategies that were used. Interviewing in a *directive* manner also appears as a way to provide respondents with retrieval cues, although survey methodologists are usually concerned with the potential to bias respondent reports.

Detailed Interviewing is so named because the behaviors assigned to this factor are those that focus the attention of the respondent on more detailed types of information, whether it be data elements within a spell (i.e., *top-down probe*), behavioral frequencies within a reference period (i.e., *how many*), or a particular calendar year (i.e., *forced year*). As these behaviors are also ones that more often populate the scripts in Q-list interviews (consider also *scripted distancing*), the inclusion within this factor of behaviors indicating the occurrence of violations of standardized interviewing (i.e., *significant change in question wording* and *wrong skip*), and of behaviors indicating that respondents were asked questions that failed to provide additional substantial information (*nothing new*), is possibly a function of those Q-list interviews that require a larger number of questions (through skip patterns) and thus provide more opportunities for interviewing mistakes and for the asking of nonapplicable questions. Negative loadings are also observed for *example* and *free year*, which are behaviors that are nearly exclusive to EHC interviews.

The factor **Cognitive Difficulty** includes *request for clarification* and *does not meet question objectives*, which are behaviors that indicate that the respondent is experiencing cognitive problems with the questions. In addition, *interviewer seeks clarification* and *interviewer clarifies* are likely to be interviewer reactions to the perceived cognitive difficulty of respondents: in the one case interviewers are seeking additional information that the respondent was not able to provide, in the other case interviewers are trying to clarify uncertainty that the respondent is experiencing. Providing *acceptable feedback* appears to occur more frequently among respondents who are having difficulty answering questions as a means to motivate continuing effort and participation.

Behaviors that include *interviewer* and *respondent's digressions* and *laughter* loaded on the **Rapport** factor, as did *unacceptable feedback*. Apparently, the affective expression that characterizes *unacceptable feedback* is more of an indication of developing rapport with respondents than other aspects of the interviewing process (see also Belli, Lepkowski, and Kabeto 2001).

5.2. Factor scores and data quality

Following Belli, Shay, and Stafford (2001), we conducted analyses on the same data quality measures derived in this earlier study on reports of income, weeks working, weeks out of the labor force, weeks unemployed, and weeks missing work due to vacation, the illness of oneself, and the illness of another. As for verbal behaviors, we limited analyses to those observed in the employment, income, and unemployment and out of the labor force, and time away domains, as these behaviors occurred during those sections of the interviews where respondents were retrieving information relevant to the data quality

Table 5. Mean differences between conditions in factor scores; Data are based on behaviors observed in employment, income, unemployment and out of the labor force, and time away domains

Factor	Condition mean (SD)		*t*-test (df)	*p*
	EHC	Q-list		
Retrieval cues	2.10 (7.60)	−2.33 (3.58)	7.69 (315)	<.0001
Detailed I'w'ing	−4.74 (2.17)	5.27 (5.91)	−22.34 (241)	<.0001
Cognitive diff	−0.33 (2.93)	0.37 (4.29)	−1.89 (337)	=.06
Rapport	−0.12 (3.13)	0.13 (4.20)	−0.66 (356)	=.51

measures. To compensate for the unequal frequency of behavior occurrences, for each factor we computed factor scores as a sum of the standard scores for each of the behaviors that loaded on the factor. An additional justification for using standard scores is that although some behaviors are observed infrequently, their actual occurrence internally may be more pronounced. For example, *spontaneous parallel* and *spontaneous sequential* behaviors may be verbally expressed far less often than they are used in the silent thought processes of respondents.

Mean factor scores for EHC and Q-list conditions, and *t*-tests to determine whether the mean differences between conditions are significant, are presented in Table 5. Whereas behaviors in the **Retrieval Cues** factor occur significantly more often in the EHC than in the Q-list condition, the behaviors in the **Detailed Interviewing** factor occur significantly more often in the Q-list than in the EHC condition. The behaviors in both the **Cognitive Difficulty** and **Rapport** factors do not differ in their level of occurrence between conditions.

The final series of analyses sought to determine whether the verbal behaviors had an effect on data quality. Three different sets of analyses were conducted. In the first set, analyses were conducted on the individual-level signed differences between experimental (1998 EHC and Q-list) and standard of comparison (1997 PSID) reports to assess the potential impact of verbal behaviors on encouraging over- or under-reporting. In the second set, analyses were conducted on the individual-level absolute value differences between experimental and standard of comparison reports to determine associations among verbal behaviors and the overall error in reports. Analyses of signed and absolute differences also examined potential experimental condition (EHC, Q-list) by verbal behavior interaction effects to determine whether the association between verbal behavior and data quality was dependent on whether EHC or Q-list interviewing had taken place. In the third set of analyses, the effect of verbal behaviors is assessed on the correlations between experimental and standard of comparison reports. As measures of data quality, correlations provide an indication of the strength of the relationship between experimental and standard of comparison reports. As these analyses on correlations are condition specific, only an indirect assessment of interaction effects is possible.

5.2.1. Analyses of signed differences

We tested separate models that examined main effects for each of the four verbal behavior factor scores (**Retrieval Cues**, **Detailed Interviewing**, **Cognitive Difficulty**, and **Rapport**) on each of the signed difference data quality measures (income, working, out of labor force, unemployment, vacation, self-ill, and other-ill). In each of these main effects models, a data quality measure was regressed on a verbal behavior factor score controlling for standard of comparison reports (e.g., when testing income, the standard of comparison report on amount of income) and the length of interview; these control variables had been found to be associated with the amount of discrepancy between experimental and standard of comparison reports. As in the case of each of the main effects models we were interested in determining the association between a verbal behavior and data quality, data from the EHC and Q-list interviews were combined. To examine potential differences in EHC and Q-list conditions, interaction models were also tested, which, in addition to including the same variables as the main effects models, included a condition (EHC, Q-list) by a verbal behavior factor score interaction term, as well as a dummy coded term for condition as a control.

Results from these analyses are reported in Column 2 of Table 6, including statistics of degrees of freedom, regression coefficients, and associated standard errors. Holm's (1979) sequentially rejective Bonferroni test procedure was implemented, at $\alpha = .05$, to control for Type I errors with significance tests on those regression coefficients obtained for each verbal behavior factor. In the main effects analyses, results show that the greater implementation of **Retrieval Cues** among interviewers and respondents is associated with an overreporting, in the experimental interviews relative to the standard of comparison, of the number of weeks worked, counteracted by an underreporting of the number of weeks out of the labor force. In addition, the greater prevalence of **Detailed Interviewing** behaviors is associated with overreporting of weeks on vacation, and the larger number of behaviors indicative of **Cognitive Difficulty** is associated with overreporting of number of weeks working and number of weeks missing work due to the illness of another. There are no significant interaction effects.

5.2.2. Analyses of absolute differences

The same models implemented in analyses of signed differences were tested using absolute differences as measures of data quality. Results are reported in Column 3 of Table 6, with the significance tests of regression coefficients being controlled for Type I errors, at $\alpha = .05$, using Holm's (1979) sequentially rejective Bonferroni test procedure. In analyses of main effects, the greater prevalence of **Retrieval Cues** is associated in a reduction in the amount of difference between experimental and standard of comparison reports of weeks missing work due to self-illness. In other words, **Retrieval Cues** are associated with less error in experimental reports for this variable. All of the remaining main effects indicate that the greater prevalence of particular behaviors is associated with increased error. Specifically, a larger number of behaviors indicating **Cognitive Difficulty** are associated with an increase in absolute differences for reports of weeks on vacation and weeks missing work due to the illness of another, and a larger number of **Rapport** behaviors are associated with higher absolute difference scores for weeks out of the labor force.

In addition to main effects, two regression coefficients demonstrate significant interaction effects. Specifically, condition and **Retrieval Cues**, and condition and **Cognitive Difficulty** behaviors, interact on reports of vacation weeks. To examine these interaction effects more thoroughly, follow-up main effects models were conducted

Table 6. Regression analyses on signed and absolute difference data quality measures. Both main and interaction effect models include standard of comparison and interview length as control variables. Interaction effect models test interaction of condition with behavior factor scores, and include terms for condition and behavior

1. Behavior/outcome	2. Signed difference				3. Absolute difference			
	Main effect		Interaction		Main effect		Interaction	
	df	β (SE)	df	β (SE)	df	β (SE)	df	β (SE)
Retrieval cues/								
income	367	197 (86.1)	365	76.1 (108)	367	21.5 (79.8)	365	−65.8 (100)
working	375	**0.27 (0.10)***	373	0.05 (0.12)	375	0.07 (0.10)	373	−0.10 (0.12)
out of labor	375	**− 0.27 (0.10)***	373	0.03 (0.12)	375	0.06 (0.10)	373	−0.07 (0.12)
unemployment	375	0.02 (0.04)	373	−0.07 (0.05)	375	−0.01 (0.04)	373	−0.04 (0.05)
vacation	375	0.02 (0.02)	373	0.02 (0.02)	375	0.01 (0.02)	373	**−0.06 (0.02)***
self-ill	375	−0.00 (0.02)	373	−0.00 (0.02)	375	**− 0.05 (0.01)***	373	0.02 (0.02)
other-ill	375	−0.01 (0.02)	373	−0.03 (0.03)	375	0.01 (0.02)	373	−0.03 (0.03)
Detailed								
income	367	−0.15 (81.9)	365	−0.91 (175)	367	76.6 (75.2)	365	341 (160)
working	375	0.11 (0.09)	373	0.17 (0.19)	375	0.02 (0.09)	373	0.25 (0.19)
out of labor	375	−0.15 (0.09)	373	−0.04 (0.18)	375	−0.02 (0.09)	373	0.18 (0.19)
unemployment	375	0.01 (0.04)	373	−0.09 (0.08)	375	0.04 (0.04)	373	0.05 (0.08)
vacation	375	**0.06 (0.02)***	373	−0.04 (0.04)	375	0.02 (0.02)	373	0.01 (0.03)
self-ill	375	−0.01 (0.02)	373	0.00 (0.04)	375	−0.01 (0.02)	373	−0.02 (0.03)
other-ill	375	0.04 (0.02)	373	−0.02 (0.04)	375	0.03 (0.02)	373	−0.01 (0.04)
Cognitive diff.								
income	367	z314 (164)	365	345 (148)	367	120 (151)	365	−22.1 (137)
working	375	**0.46 (0.17)***	373	0.24 (0.16)	375	0.15 (0.17)	373	−0.29 (0.16)
out of labor	375	−0.40 (0.17)	373	0.18 (0.16)	375	0.18 (0.17)	373	−0.25 (0.16)
unemployment	375	−0.03 (0.07)	373	−0.04 (0.07)	375	−0.10 (0.07)	373	−0.02 (0.07)
vacation	375	0.05 (0.03)	373	−0.03 (0.03)	375	**0.09 (0.03)***	373	**− 0.09 (0.03)***
self-ill	375	−0.03 (0.03)	373	−0.00 (0.03)	375	−0.00 (0.03)	373	−0.02 (0.03)
other-ill	375	**0.11 (0.04)***	373	−0.06 (0.04)	375	**0.11 (0.04)***	373	−0.06 (0.04)

Table 6. Continued

1. Behavior/outcome	2. Signed difference				3. Absolute difference			
	Main effect		Interaction		Main effect		Interaction	
	df	β (SE)	df	β (SE)	df	β (SE)	df	β (SE)
Rapport								
income	367	94.0 (142)	365	132 (140)	367	−72.1 (130)	365	−99.9 (128)
working	375	0.29 (0.16)	373	−0.28 (0.15)	375	0.36 (0.15)	373	−0.24 (0.15)
out of labor	375	−0.31 (0.16)	373	0.35 (0.15)	375	**0.41 (0.15)***	373	−0.07 (0.15)
unemployment	375	0.02 (0.07)	373	−0.01 (0.07)	375	−0.04 (0.07)	373	−0.01 (0.06)
vacation	375	0.02 (0.03)	373	−0.04 (0.03)	375	0.04 (0.03)	373	−0.02 (0.03)
self-ill	375	−0.01 (0.03)	373	−0.02 (0.03)	375	−0.03 (0.03)	373	−0.00 (0.03)
other-ill	375	0.05 (0.04)	373	−0.04 (0.04)	375	0.05 (0.04)	373	−0.03 (0.04)

*Significant at alpha = .05 adjusted using Holm's sequentially rejective multiple Bonferroni test procedure within rows and columns.

separately for EHC and Q-list conditions. The Q-list condition shows a significant positive association between **Retrieval Cues** and absolute error in reports of vacation weeks, $\beta = 0.116$, SE $= .049$, $t(172) = 2.35$, $p = .02$; the EHC condition, on the other hand, demonstrates a nonsignificant negative association, $\beta = -0.009$, SE $= 0.015$, $t(199) = -0.61$, $p = .55$. A similar pattern is found for **Cognitive Difficulty** behaviors, in which there is a significant positive association in the Q-list condition, $\beta = 0.150$, SE $= 0.042$, $t(172) = 3.60$, $p < .001$, but a nonsignificant negative association in the EHC condition, $\beta = -0.008$, SE $= 0.037$, $t(199) = -0.21$, $p = .84$. In sum, in the Q-list condition, for reports of weeks on vacation, there is an increase in error when the prevalence of **Retrieval Cues** and **Cognitive Difficulty** behaviors is higher, but there is no association between these behaviors and measures of data quality in the EHC condition.

5.2.3. Analyses of correlations

The measure of data quality in these analyses is the correlation between experimental and standard of comparison reports for each outcome measure (income, working, out of labor force, unemployment, vacation, self-ill, and other-ill). Because interaction effects could not be directly inferred, the EHC and Q-list conditions are treated separately in order to descriptively assess whether verbal behavior patterns affect the conditions differentially. Hence, within each condition the median of the verbal behavior factor scores for each factor (**Retrieval Cues**, **Detailed Interviewing**, **Cognitive Difficulty**, and **Rapport**) was computed, and Pearson correlations were computed separately for those interviews in which factor scores were higher than the median (high interviews), and for those in which factor scores were lower (low interviews). z-tests were next conducted in a comparison of high and low interview conditions to determine whether their respective correlation coefficients significantly differed in strength.

During an initial evaluation of analyses, it became apparent that significance testing would be inappropriate if the weaker correlation had been observed in those interviews that consisted of less variation in standard of comparison (or experimental) reports. Consider, for example, results that show that in standard of comparison reports, not only is the mean level of weeks working significantly higher (high $M = 47.2$; low $M = 26.3$), and the mean level of weeks out of labor force lower (high $M = 3.7$; low $M = 24.6$), in EHC interviews with high prevalence of **Retrieval Cues** than in low interviews, $t(149) = 7.40$, $p < .0001$, and $t(142) = -7.52$, $p < .0001$, respectively, but the variation in both weeks working (high SD $= 12.5$; low SD $= 25.6$) and out of labor force (high SD $= 11.4$; low SD $= 25.7$) is lower in the high compared to the low interviews: Folded $F(102, 99) = 4.19$, $p < .0001$, and Folded $F(102,99) = 5.09$, $p < .0001$, respectively. A reasonable explanation of these results is that in situations in which there are more work weeks to retrieve, more **Retrieval Cues** are used to aid in the retrieval process. Moreover, as **Retrieval Cues** are used in interviews that are populated by cases dominated by having many weeks of work (and few out of labor force weeks), the variation in the number of weeks working (and out of labor force) is truncated in cases in which **Retrieval Cues** are likely to occur. Hence, although the correlation coefficient is weaker in high interviews than in low interviews, this difference is not informative as correlation coefficients are reduced when the variation in the distributions is truncated, which is the case in the high condition with the standard of comparison (and, by the way, the experimental) distributions for weeks working and weeks out of the labor force.

Accordingly, z-tests were conducted only for those comparisons between high and low interviews in which there are no significant differences in levels of variation of either standard of comparison or experimental reports, as long as, of course, the variation differences are in the same direction as the correlation coefficients themselves. The results of the correlation analyses are presented in Table 7. To control for Type I errors, the significance of the z-tests, at $\alpha = .05$, is adjusted using Holm's (1979) sequentially rejective Bonferroni test procedure.

Table 7. Correlations between experimental and standard of comparison reports for income, unemployment, self-illness, and other-illness, as a function of a median split for factor scores conducted separately for EHC and Q-list conditions

	EHC			Q-List		
Factor/outcome	High	Low	z	High	Low	z
Retrieval cues/						
income	.932	.873	2.25*	.790	.979	−6.74*
working	.575	.868	NI	.784	.933	NI
out of labor	.461	.889	NI	.758	.918	NI
unemployment	.807	.610	2.81*	.058	.883	−8.63*
vacation	.587	.487	NI	.028	.835	−7.63*
self-ill	.774	.517	NI	.320	.000	2.16*
other-ill	.673	.271	NI	.109	.772	−5.94*
Detailed						
income	.794	.984	−9.28*	.821	.985	−8.52*
working	.638	.908	NI	.588	.919	NI
out of labor	.698	.895	NI	.648	.883	NI
unemployment	.120	.879	NI	.186	.622	−3.51*
vacation	.426	.648	−2.22*	.122	.789	−6.15*
self-ill	.801	.625	NI	.109	.085	0.16
other-ill	.654	.276	NI	.106	.802	−6.42*
Cognitive difficulty						
income	.935	.855	2.90*	.955	.849	4.14*
working	.613	.863	NI	.716	.954	NI
out of labor	.504	.622	NI	.658	.943	NI
unemployment	.794	.885	NI	.304	.217	0.62
vacation	.499	.587	−0.87	.414	.294	NI
self-ill	.457	.909	NI	.078	.234	−1.05
other-ill	.254	.780	NI	.100	.953	−11.48*
Rapport						
income	.882	.904	−0.74	.823	.974	−6.57*
working	.817	.858	NI	.875	.931	−2.00
out of labor	.767	.887	NI	.906	.893	0.55
unemployment	.850	.091	NI	.045	.554	−3.76*
vacation	.550	.542	NI	.133	.722	−5.06*
self-ill	.720	.662	NI	.105	.203	−0.65
other-ill	.740	.232	NI	.100	.973	−13.35*

NI−not informative due to unequal variances in distributions.

*Significant at alpha = .05 adjusted using Holm's sequentially rejective multiple Bonferroni test procedure within rows and columns.

Results indicate modest support for the notion that the use of **Retrieval Cues** assists in the remembering of events in the EHC condition, as Pearson correlations between experimental and standard of comparison reports are significantly stronger in high interviews for income and weeks unemployed than in low interviews. As for the Q-list condition, a higher prevalence of **Retrieval Cues**, with the exception of reports of weeks missing work due to self-illness, leads to significantly weaker correlations for all informative comparisons. Results for **Detailed Interviewing** in both EHC and Q-list conditions are consistent in demonstrating significantly weaker correlations among a number of data quality measures in the high interviews in comparison to the low ones. The correlation coefficients between high and low prevalence interviews in **Cognitive Difficulty** behaviors show inconsistent results. In both the EHC and Q-list conditions, reports for income show stronger correlations when the prevalence of **Cognitive Difficulty** behaviors is higher; yet for reports of weeks missing work due to another's illness, a stronger correlation appears in interviews that are lower in **Cognitive Difficulty** behaviors in the Q-list condition. Finally, **Rapport** behaviors consistently show significantly weaker correlations in the high interviews in the Q-list condition. In the EHC condition, the differences in the levels of **Rapport** between low and high interviews are mostly not informative and inconsistent in their direction.

6. Summary and Concluding Remarks

The analyses reported in this article were designed to assess whether interviewers' and respondents' differential use of retrieval cues, associated with parallel and sequential retrieval, and differential use of conversational processes, associated with expressions of uncertainty, cognitive problems, violations of standardization, and rapport, could account for the observation that EHC interviews lead to higher quality retrospective reports than do Q-list interviews. Audiotaped interviews from both EHC and Q-list interviews were coded for retrieval and conversational behaviors, and specific types of behaviors were associated with indices on the quality of retrospective reports.

6.1. Summary and interpretation of results

6.1.1. Prevalence of behaviors

EHC interviews, as hypothesized, were marked by the more frequent use of a variety of parallel and sequential retrieval cues in the probing by interviewers and in the spontaneously used retrieval strategies of respondents. Also consistent with the more flexible style of interviewing that EHCs promote, in comparison to Q-list interviews, EHC interviews were noted as encouraging a more open approach in allowing respondents to temporally order which question objectives to report, and EHC interviewers engaged in a larger number of attempts to ascertain the adequacy of respondent reports. Despite the greater flexibility engendered by EHC interviews, in comparison to Q-list methods there was no evidence to indicate a greater degree of interviewer variance in response quality.

6.1.2. Relationship of verbal behaviors to data quality

Results of analyses examining associations between sets of verbal behaviors (ascertained through a principle components analysis) and data quality measures provide several insights into the potential influence of behaviors on the accuracy of retrospective survey reports in both EHC and Q-list interviewing methodologies. An important finding is that a set of behaviors marked by the use of retrieval cues interacts with EHC and Q-list conditions in their association with data quality. Although not overwhelming, there is evidence that supports the hypothesis that a more extensive use of retrieval cues improves the quality of retrospective reports in EHC interviews, but is detrimental to Q-list ones. Apparently, a pattern of verbal behaviors that encourages using retrieval cues within the structure of autobiographical memory, although beneficial for EHC interviews, is not beneficial for Q-list ones. As some of the interviewer retrieval probing behaviors have low prevalence in EHC interviews, particularly those dealing with parallel retrieval, results might had been stronger if interviewers were able to implement retrieval probes more often. One aim of additional work with EHC interviewing, then, would be to implement design innovations and interviewer training that would encourage the greater use of parallel retrieval probes.

A set of behaviors that indicate both violations of standardization and the lack of a flexible or open approach are consistently detrimental to data quality in both EHC and Q-list conditions. However, a simple interpretation of results for this set of behaviors is misleading because the behaviors that contribute to differences in data quality were at times different ones, depending on condition. As for both EHC and Q-list interviews, it appears that forcing a retrieval order on respondents (either moving chronologically forward or backward) is detrimental to data quality, whereas better data quality is observed when respondents are allowed to choose whatever retrieval order they prefer. This interpretation is consistent with findings that have shown that allowing individuals to choose retrieval order leads to more accurate remembering than forcing a forward or backward chronological retrieval order (Loftus and Fathi 1985; Jobe et al. 1990). Another possible explanation of an association between these behaviors and response quality in the Q-list condition is that more of these detrimental behaviors may simply occur as an artifact of more complicated interviews, in which more questions are asked respondents. In turn, more complicated interviewing situations will lead to poorer data quality than less complicated ones. As another possibility, as several behaviors that populate this set are ones in which prescriptions of standardized interviewing are violated, such as making changes in question wording that alter the intended meaning or following an incorrect skip pattern, increases in these behaviors may have a detrimental effect on data quality, as often predicted by survey methodologists (Beatty 1995; Fowler and Cannell 1996; Oksenberg et al. 1991) but until now without empirical support (Belli and Lepkowski 1996; Dykema, Lepkowski, and Blixt 1997).

Although problems with cognitive processes had mixed influences on data quality, a behavior by condition interaction with a set of behaviors that reflect cognitive difficulty indicate that the occurrence of cognitive problems is more consistent in leading to poorer quality retrospective reports in Q-list interviews than in EHC ones. Similarly, the presence of rapport behaviors shows a greater influence of being detrimental to retrospective reports in Q-list than in EHC interviews. These results partly replicate those of Belli, Lepkowski, and Kabeto (2001), who found, in retrospective reports for doctor's office visits in Q-list interviews, that behaviors reflecting cognitive difficulty are associated with poorer data quality. The conversational flexibility of EHC interviews likely can offset the potentially deleterious consequences of cognitive difficulty and rapport. With regard to cognitive difficulty, interviewers are more likely to seek clarification of cognitive uncertainties in EHC than in Q-list interviews, and there is evidence that a more unconstrained, conversational approach toward resolving problems of meaning is beneficial (Schober and Conrad 1997). As for rapport, the relationship-building between participants that rapport represents may flow more naturally in a flexible, conversational exchange, whereas in standardized Q-list interviewing, rapport may distract respondents from their task of answering scripted questions (Dijkstra 1987).

6.2. Limitations

There are two main limitations to our results. The first one centers on inherent problems with verbal behavior coding data. Although such coding is a valuable technique to gain insights into cognitive and conversational processes, limitations arise in there not being a clear determination regarding what distinguishes one utterance from another, with researchers and coders possibly imposing their own judgments concerning which behaviors are of interest (Ongena 2002). Gaining reliability in code assignments among more than one coder allays these concerns, but it never completely eliminates them. In addition, verbal behavior coding can only provide insights into cognitive processes that are revealed by overt speech. Verbal or other cognitive processes that occur silently are beyond observation, although cognitive interviewing techniques have been able to reveal the importance that ordinarily silent speech plays in the cognitive processing of respondents (Forsyth and Lessler 1991; Willis, Royston, and Bercini 1991). Of course, a verbal behavior coding overcomes limitation of cognitive interviewing in that it can reveal cognitive and conversational processes that may uniquely occur within the actual context of asking survey questions (Fowler and Cannell 1996).

The second limitation centers on the reliability of data from a standard of comparison that, like the data for the experimental conditions, has error properties associated with its origins in verbal reports. Because the retention interval between the occurrence of events and the standard of comparison reports is shorter than the retention interval for the experimental reports, the assumption that the error properties with the standard of comparison ought to be less severe, other things being equal, than those that exist with the experimental reports, is reasonable. Nevertheless, there are no clear specifications regarding what the exact characteristics of the error structures within these data sets are, and as one anonymous reviewer noted, inferences regarding differences in data quality between conditions would be more convincing if true score measures were obtained, perhaps by asking participants to report directly from their 1996 tax forms for income information at the end of the administration of the experimental interviews. In hindsight, seeking true score measures could be a valuable exercise, and one that should be considered in future research. However, because of the intrusive nature of these requests, care would need to be taken not to alienate respondents (especially if they are participants in a panel survey), additional costs will have to be covered, and because not all participants will honor these requests, problems with selection bias would have to be examined for potential threats to validity.

6.3. *Conclusion*

The results suggest that there exists a differential effect of verbal behaviors on data quality depending on interviewing methodology. The standardization of Q-list interviews appears to be one in which 1) retrieval cues are used infrequently, and when used, they may encourage a style of narrative remembering that interferes with the restricted answer formats that are offered, 2) violations in standardization such as asking questions that are at variance with scripted wording may harm data quality, 3) the cognitive difficulty experienced by respondents is detrimental to the quality of retrospective reports (Belli and Lepkowski 1996; Belli, Lepkowski, and Kabeto 2001; Fowler 1992), and 4) behaviors that are outside the context of standardization per se, especially ones in which interviewers and respondents develop a personal relationship with one another, can lead to poorer data quality (Beatty 1995; Dijkstra 1987; Williams 1968). In contrast, the conversationally flexible style of interaction between interviewers and respondents that characterizes EHC interviews appears to be one in which 1) a more extensive use of retrieval cues enhances the quality of retrospective reports (Belli 1998), 2) permitting respondents to use whichever chronological direction of retrieval they choose is beneficial to data quality (Loftus and Fathi 1985; Jobe et al. 1990), 3) the detrimental consequences of respondents' cognitive difficulty may be attenuated by interviewers conversationally seeking clarification (Schober and Conrad 1997), and 4) the development of rapport between interviewers and respondents does not detract from the beneficial impact of conversational interviewing.

The contention of survey methodologists that adherence to strict standardization and maintenance of a task-oriented relationship produces the best quality data with Q-list interviewing methods is supported by the results. As for EHC interviewing, a flexible conversational style that implements retrieval cues and an open approach to respondent reporting produces the best quality data. In choosing between these methods, the results favor EHC methods overall. The quality of retrospective reports has been found to be better with EHC methods (Belli, Shay, and Stafford 2001; Yoshihama et al. 2003). In addition, these benefits have not proved to come at an increased cost in interviewing time or interviewer variability, although more research to examine these factors is needed to provide firm evidence that increased costs are not involved. In conclusion, the results provide encouraging support for EHC interviewing becoming one day the more preferred approach in collecting retrospective report data.

7. References

Bailar, B.A., Bailey, L., and Stevens, J. (1977). Measures of Interviewer Bias and Variance. Journal of Marketing Research, 14, 337–343.

Barsalou, L.W. (1988). The Content and Organization of Autobiographical Memories. In Remembering Reconsidered: Ecological and Traditional Approaches to the Study of Memory, U. Neisser and E. Winograd (eds). New York: Cambridge University Press, 193–243.

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

Belli, R.F., Lepkowski, J.M., and Kabeto, M.U. (2001). The Respective Roles of Cognitive Processing Difficulty and Conversational Rapport on the Accuracy of Retrospective Reports of Doctor's Office Visits. In Seventh Conference on Health Survey Research Methods, Marcie L. Cynamon and Richard A. Kulka (eds). Hyattsville, MD: U.S. Government Printing Office, (DHHS Publication No. (PHS) 01-1013), 197–203.

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.

Cannell, C.F., Lawson, S.A., and Hausser, D.L. (1975). A Technique for Evaluating Interviewer Performance. Ann Arbor: The University of Michigan.

Cannell, C.F. and Oksenberg, L. (1988). Observation of Behavior in Telephone Interviews. In Telephone Survey Methodology, R.M. Groves, P. Biemer, L. Lyberg, J. Massey, W. Nicholls, and J. Waksberg (eds). New York: Wiley.

Conrad, F.G. and Schober, M.F. (2000). Clarifying Question Meaning in a Household Telephone Survey. Public Opinion Quarterly, 64, 1–28.

Conway, M.A. (1996). Autobiographical Knowledge and Autobiographical Memories. In Remembering Our Past: Studies in Autobiographical Memory, D.C. Rubin (ed.). New York: Cambridge University Press, 67–93.

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

Dijkstra, W. (2002). Transcribing, Coding, and Analyzing Verbal Interactions in Survey Interviews. In Standardization and Tacit Knowledge, D.W. Maynard, H. Houtkoop-Steenstra, N.C. Schaeffer, and J. van der Zouwen (eds). New York: Wiley, 401–425.

Dykema, J., Lepkowski, J.M., and Blixt, S. (1997). The Effect of Interviewer and Respondent Behavior on Data Quality: Analysis of Interaction Coding in a Validation Study. In Survey Measurement and Process Quality, L. Lyberg, P. Biemer, M. Collins, E. DeLeeuw, C. Dippo, N. Schwarz, and D. Trewin (eds). New York: Wiley, 287–310.

Forsyth, B.H. and Lessler, J.T. (1991). Cognitive Laboratory Methods: A Taxonomy. In Measurment Errors in Surveys, P.B. Biemer, R.M. Groves, L.E. Lyberg, N.A. Mathiowetz, and S. Sudman (eds). New York: John Wiley, 393–418.

Fowler, F.J. (1992). How Unclear Terms Affect Survey Data. Public Opinion Quarterly, 56, 218–231.

Fowler, F.J. and Cannell, C.F. (1996). Using Behavioral Coding to Identify Cognitive Problems with Survey Questions. In Answering Questions, S. Schwarz and S. Sudman (eds). San Francisco: Jossey-Bass, 15–36.

Groves, R.M. (1989). Survey Errors and Survey Costs. New York: Wiley.

Holm, S. (1979). A Simple Sequentially Rejective Multiple Test Procedure. Scandinavian Journal of Statistics, 6, 65–70.

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

Jobe, J.B., White, A.A., Kelley, C.L., Mingay, D.J., Sanchez, M.J., and Loftus, E.F. (1990). Recall Strategies and Memory for Health-Care Visits. The Milbank Quarterly, 68, 171–189.

Kish, L. (1962). Studies of Interviewer Variance for Attitudinal Variables. Journal of the American Statistical Association, 57, 92–115.

Loftus, E.F. and Fathi, D.C. (1985). Retrieving Multiple Autobiographical Memories. Social Cognition, 3, 280–295.

Mangione, T.W., Fowler, F.J., and Lewis, T.A. (1992). Question Characteristics and Interviewer Effects. Journal of Official Statistics, 8, 293–307.

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

Ongena, Y. (2002). Methods of Behavior Coding for Survey Interviews. Paper presented at the Methods for Studying Interaction Workshop, University of Wisconsin, Madison, April.

Presser, S. and Blair, J. (1994). Survey Pretesting: Do Different Methods Produce Different Results? In Sociological Methodology, P.V. Marsden (ed.). Washington, DC: American Sociological Association.

Schaeffer, N.C., Maynard, D.W., and Cradock, R.M. (1993). Negotiating Certainty: Uncertainty Proposals and Their Disposal in Standardized Interviewing. University of Wisconsin–Madison, Center for Demography and Ecology, Working Paper, 93-25.

Schober, M.F. and Conrad, F.G. (1997). Does Conversational Interviewing Reduce Survey Measurement Error? Public Opinion Quarterly, 61, 576–602.

Van der Vaart, W. (2002). The Time-Line: The Effects of an Experimental Aided Recall Technique in a Real Life Survey. Paper presented at the International Conference on Questionnaire Development, Testing, and Evaluation Methods, Charleston, South Carolina, November.

Van der Zouwen, J. and Dijkstra, W. (2002). Testing Questionnaires Using Interaction Coding. In Standardization and Tacit Knowledge, D.W. Maynard, H. Houtkoop-Steenstra, N.C. Schaeffer, and J. van der Zouwen (eds). New York: Wiley, 427–447.

Williams, J.A. (1968). Interviewer Role Performance: A Further Note on Bias in the Information Interview. Public Opinion Quarterly, 32, 287–294.

Willis, G.B., Royston, P., and Bercini, D. (1991). The Use of Verbal Report Methods in the Development and Testing of Survey Questionnaires. Applied Cognitive Psychology, 5, 251–267.

Yoshihama, M., Gillespie, B., Hammock, A.C., Belli, R.F., and Tolman, R.M. (2003). Does the Life History Calendar Method Facilitate the Recall of Domestic Violence Victimization? Forthcoming.

Received February 2003
Revised February 2004