DE GRUYTER
OPEN

Journal of Official Statistics, Vol. 32, No. 3, 2016, pp. 579–600, **http://dx.doi.org/10.1515/JOS-2016-0030**

Using Data Mining to Predict the Occurrence of Respondent Retrieval Strategies in Calendar Interviewing: The Quality of Retrospective Reports

Robert F. Belli[1], *L. Dee Miller*[2], *Tarek Al Baghal*[3], *and Leen-Kiat Soh*[4]

Determining which verbal behaviors of interviewers and respondents are dependent on one another is a complex problem that can be facilitated via data-mining approaches. Data are derived from the interviews of 153 respondents of the Panel Study of Income Dynamics (PSID) who were interviewed about their life-course histories. Behavioral sequences of interviewer-respondent interactions that were most predictive of respondents spontaneously using parallel, timing, duration, and sequential retrieval strategies in their generation of answers were examined. We also examined which behavioral sequences were predictive of retrospective reporting data quality as shown by correspondence between calendar responses with responses collected in prior waves of the PSID. The verbal behaviors of immediately preceding interviewer and respondent turns of speech were assessed in terms of their co-occurrence with each respondent retrieval strategy. Interviewers' use of parallel probes is associated with poorer data quality, whereas interviewers' use of timing and duration probes, especially in tandem, is associated with better data quality. Respondents' use of timing and duration strategies is also associated with better data quality and both strategies are facilitated by interviewer timing probes. Data mining alongside regression techniques is valuable to examine which interviewer-respondent interactions will benefit data quality.

Key words: Calendar interviewing; data mining; interviewing; memory aids.

1. Introduction

In the collection of retrospective reports, calendar interviewing methods have reliably led to better data quality in comparison to conventional standardized methods, at times with only limited costs in which increases in interviewing and programming time are negligible or minimal at most (for reviews see Belli 2014; Belli and Callegaro 2009; Glasner and van der Vaart 2009). In calendar interviews, instead of having questions written in advance as in conventional standardized interviewing, interviewers develop queries to satisfy questionnaire objectives that are largely visually displayed by timelines within various domains (see, for example, Balán et al. 1969; Freedman et al. 1988). Each timeline is constructed with

[1] University of Nebraska, Department of Psychology, Lincoln, NE 68588-0308, U.S.A. Email: bbelli2@unl.edu
[2] University of Nebraska, 2343 Stone Creek Loop South, Lincoln, NE 68512, U.S.A. Email: ldeemiller@gmail.com
[3] University of Essex, ISER, Colchester, UK CO4 3SQ. Email: talbaghal@gmail.com
[4] University of Nebraska, 122E Avery Hall, Lincoln, NE 68588-0115, U.S.A. Email: lksoh@cse.unl.edu

Acknowledgments: This article is based upon work supported by the National Science Foundation under Grant No. 1132015. Any opinions, findings, and conclusions or recommendations expressed in this article are those of the authors and do not necessarily reflect the views of the National Science Foundation.

a specified unit of analysis (e.g., week, month, or year) and reference period (e.g., one year, ten years, or from birth to the present), and they are aligned with calendar time depending on their unit of analysis. Within each timeline, queries by interviewers will seek to get respondents to report periods of stability and points of transition, such as being employed with one employer for a period of time and then transitioning to another employer at another period of time. A domain represents a topic of interest, such as information on residential, partnering, parenting, labor, and health histories, and each domain may consist of one to several timelines. For example, when collecting labor histories, separate timelines may be devoted to employment and unemployment, respectively.

The improvements in data quality with calendar interviewing methods have been examined both theoretically and empirically within the context of the structure of autobiographical memory (Belli 1998; Belli et al. 2007; Bilgen and Belli 2010). Specifically, calendar methods have been shown to encourage the use of verbal retrieval behaviors in both interviewers and respondents that, in comparison to conventional questionnaires, are associated with better data quality for retrospective reports of life-course labor histories (Belli et al. 2007), especially for respondents who have experienced complicated pasts (Belli et al. 2013). Further, these behaviors align with the structure of autobiographical memory (Belli et al. 2004; Bilgen and Belli 2010).

Although calendar methods have produced encouraging results, as noted by Belli et al. (2013), these results are limited because they do not examine the communicative interactions between interviewers and respondents directly. In this article, to overcome this limitation, we examine those series of communicative interactions that are most likely to lead to respondents' use of retrieval strategies. We focus on respondent retrieval strategies as the outcome of interviewer-respondent interactions because we believe that their use is tied most directly to the successful remembering of past events. As for interviewer retrieval probing, our expectation is that the use of these probes will promote the use of retrieval strategies on the part of respondents, a result that we expect to confirm via our interactional analyses.

In terms of the structure of autobiographical memory, we examine those retrieval strategies *consisting of parallel and sequential cues* (Belli 1998; Belli and Callegaro 2009; Belli et al. 2013). With *parallel* retrieval strategies, respondents cue themselves by remembering a contemporaneous event from a different life domain as an apparent attempt to more fully reconstruct the past. An example of parallel cuing would occur if a respondent is asked about when a job ended, and they spontaneously remember the birth of a child when answering this query. With *sequential* retrieval strategies, respondents seek to order what happened earlier and later in time within the same domain by seeking to remember the time location of the beginning and ending of events, the duration of events, and/or what event occurred earlier or later. An example of sequential retrieval would occur if the respondent remembered that working as a librarian at a university immediately followed working as an office worker for a private company.

Hence, we are concerned with two main issues. First, we seek to determine which series of verbal exchanges between interviewers and respondents in calendar interviews are more likely to lead to respondent retrieval strategies. Earlier research examining interviewer and respondent verbal exchanges with conventional questionnaires has demonstrated the challenges of these approaches. Brenner (1982) was interested in determining via tree structures those combinations of behaviors that followed earlier specific behaviors, such as what follows from interviewers asking questions as written versus when interviewers altered questions. Although such modeling could be applied in reverse, so that the tree structure from a later behavior could be propagated forward in time, the hand calculation of these tree structures is cumbersome. In addition, Brenner (1982) concentrated only on the occurrence of behaviors; we are also interested in determining whether the nonoccurrence of behaviors is similarly predictive of a final respondent retrieval. Adding nonoccurrence leads to further computational challenges. The only study we found that sought to examine which behaviors occurred earlier focused only on single behaviors (Dijkstra and Ongena 2006), and not on different combinations of earlier behaviors. In order to identify those series of behaviors that are predictive of the occurrence of respondent retrievals, while accounting for the occurrence and nonoccurrence of those behaviors contained in these series, we use data-mining techniques that have been developed in the field of computer science.

Second, having identified different behavioral series that are predictive of the presence of respondent retrieval behaviors, we conduct analyses to determine which, if any, of these series are predictive of retrospective reports of better data quality. Although, as noted above, Belli et al. (2013) have demonstrated that respondent retrieval strategies are associated with better data quality, their research used a confirmatory factor-analysis approach to create a single latent measure of respondent retrieval from several behaviors. This work extends that in Belli et al. (2013) by providing more focused interactional analyses. Specifically, by identifying which behavior series are predictive of better data quality and which are not, we show that behavioral interactions between interviewers and respondents lead to respondent retrieval strategies that vary in their effectiveness.

2. Data-Collection Method

Response data were collected from 313 Panel Study of Income Dynamics respondents of 45 years of age and older in 2002 (93% cooperation rate, AAPOR standard definition 1). Respondents were interviewed with a computer-assisted telephone interviewing (CATI) calendar instrument that asked for reports on residence, relationship, labor (employment and unemployment), and health lifetime histories. 297 interviews were audio recorded with respondent permission, with 291 audible tapes transcribed. Greater detail on the calendar CATI data-collection methods can be found in Belli et al. (2007) and Belli et al. (2013).

A random sample of 165 interviews was behavior coded with a scheme that comprised 30 interviewer and 29 respondent verbal behaviors. Behaviors were identified within turns of speech, a turn being defined as a transcribed uninterrupted utterance by either the interviewer or respondent. Greater detail on the behavior-coding methods, the reliability among coders, and the verbal behaviors that were identified can be found in Bilgen and Belli (2010).

3. Data Analyses and Results

3.1. Data-mining Algorithm

Our overall aim is to implement a data-mining algorithm able to isolate different series of verbal behaviors immediately preceding three turns of speech – an interviewer turn, a respondent turn, and another interviewer turn – to those respondent turns that contained one of four respondent retrieval strategies. We selected three preceding turns of speech as an attempt to come to some compromise in which either too few or too many turns of speech would be subjected to analysis. We did not want to select only the single turn of speech that immediately preceded the targeted turn as we understood that especially in calendar interviews, the behaviors of turns that had occurred earlier could have a lasting influence for a number of subsequent turns. However, we also did not want to extend our analyses too far backward, as impact would diminish as the number of intervening turns increased. With these constraints in mind, we fully understand that isolating three preceding turns is based on more subjective than empirical criteria, and that future work may wish to examine more turns.

For partly empirical (e.g., Belli et al. 2013) and partly theoretical reasons (e.g., Belli 1998), the four respondent retrieval strategies we examined were parallel, timing, duration, and sequential behaviors. These four behaviors exhaust what our empirical work in behavior coding has discovered as comprising both parallel and sequential retrieval strategies, which are those respondent retrieval strategies that have been theoretically hypothesized and empirically shown to be associated with better data quality in calendar interviews. All occurred spontaneously in respondents' verbal behavior, that is, the behaviors were not a direct reflection of a query made by an interviewer. A *parallel* retrieval strategy occurred when a respondent spontaneously remembered a contemporaneous event from a life domain that was different than the one being queried. A *timing* retrieval strategy was present when a respondent spontaneously indicated a beginning or ending location in time of a reported event. A *duration* retrieval strategy consisted of spontaneous reports of the length in time of events. Finally, a *sequential* retrieval strategy occurred when a respondent spontaneously reported an event that occurred earlier or later than one which had already been identified within the same domain.

Of the 35,291 transcribed respondent turns of speech from the 165 interviews, 1,744 were identified as including a respondent parallel retrieval, and 2,821, 1,191, and 765 included timing, duration, and sequential behaviors, respectively. In order to apply the data-mining algorithm, in addition to turns which included respondent retrieval strategies, we had to simultaneously analyze turns in separate tests for each behavior that did not consist of any respondent retrievals in order to find series in the preceding turns that were diagnostic of each of the behaviors in the targeted turns. As nonretrieval turns are considerably more numerous than those turns that contain a respondent retrieval, problems associated with imbalance arise. Data-mining algorithms tend to assume relatively balanced distributions (He and Garcia 2009) as imbalanced data sets reduce the predictive power of these algorithms (Weiss and Provost 2001).

To achieve balanced distributions, we kept all turns that did include a respondent retrieval behavior and randomly sampled, for each behavior separately, an equal number of respondent turns that did not include the targeted behavior. Hence, we conducted four separate analyses in which a total of 3,488 turns of speech were examined for parallel retrieval behaviors in one analysis, and 5,642, 2,382, and 1,530 turns were examined for timing, duration, and sequential retrieval behaviors in each of three analyses, respectively. We adopted a decision-tree data-mining algorithm called C4.5 (Witten et al. 2011) and separately applied it to each of the four respondent retrieval behaviors in their respective analyses. This algorithm was used to discover what behavioral series in the prior three turns are most predictive of the state for the respondent retrievals in the targeted fourth turn.

The decision-tree algorithm "grows" multiple-behavior series using a top-down approach with two heuristics. First, the algorithm applies a heuristic to choose the behavior that most improves the predictive power for the series (*behavior-select heuristic*) from all of the 30 interviewer and 29 respondent verbal behaviors that could potentially appear in the selected preceding three turns. The behavior chosen by this heuristic is added to the tree as an internal node. More on this behavior-select heuristic (based on information gain) will be discussed later in this section.

After the application of the behavior-select heuristic, the algorithm divides the turns into two groups based on the occurrence or nonoccurrence of the chosen behavior. The first group contains the turns with occurrence of that behavior, while the second group contains turns with nonoccurrence. The occurrence and nonoccurrence are added as edges under the node in the tree.

Now, the algorithm uses another heuristic to decide whether growing the series further would improve its predictive power (*continue-growing heuristic*). The heuristic makes this decision *separately* for each group by evaluating (1) the distribution of the respondent behavior for turns in the group and (2) the size of the group. In general terms, there are two cases where the continue-growing heuristic should decide to stop:

1. The heuristic should stop when the group has a sufficiently homogenous state for the respondent behavior (present or absent) – the series has already mastered the group and growing the series further will not improve predictive power.
2. The heuristic should stop when the group is too small, since growing the series on outliers could actually hurt overall predictive power; that is, it could lead to overfitting.

In both cases, the series is ready to make a final prediction (on the group) since predictive power is unlikely to be improved. The final prediction for the respondent behavior is set to the majority respondent-behavior state for the turns in the group. This final prediction is then added as a node to the decision tree.

Alternatively, the heuristic should continue to grow the series to try and improve its predictive power. To this end, the decision-tree algorithm restarts, running the above process again using only the turns in the group. The new behaviors are added as additional internal nodes connected to the previous behavior occurrence and nonoccurrence edges.

Figure 1 provides a graphical illustration for a possible decision tree. The behaviors chosen are listed inside boxes with the occurrence or nonoccurrence of these behaviors given as edges. The round nodes are the final prediction for the respondent behavior on a group. As alluded to earlier, the decision tree allows for multiple series of behaviors to be grown. These series can be discovered by tracing a path through the decision tree from the behavior at the top to a circle. As an example, the occurrence of Behavior A at any of the preceding three turns combined with occurrence of Behavior B at any of the preceding three turns leads to a targeted respondent parallel being present at the fourth turn, as indicated by a yes circle.

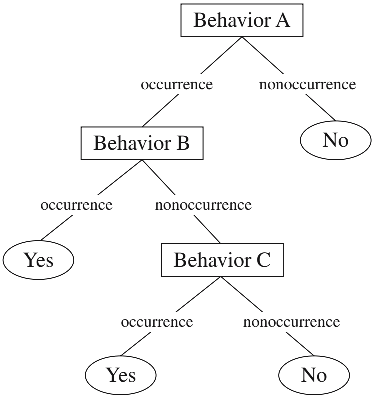

Fig. 1. Graphical example of decision tree. Behaviors are given in boxes with occurrence or nonoccurrence on the edges. The circle indicates whether or not respondent behavior is present (Yes or No). Sequences are discovered by tracing from the behavior at the top (Behavior A) to a circle. Example sequence on the above tree: occurrence of Behavior A combined with occurrence of Behavior B leads to respondent-parallel being present.

Behavior-Select Heuristic

Our behavior-select heuristic uses the same information gain as the C4.5 algorithm. We first test all of the behaviors that occur in the previous turns separately and calculate the extent to which the presence and absence of each behavior affects the distribution for the respondent retrieval in the targeted fourth turn. A completely homogenous distribution is one that only contains retrieval-present turns, or retrieval-absent turns. Hence, and using respondent parallel retrievals as an example, the greatest degree of homogeneity would result if a behavior was identified that (1) when it occurred in the previous turns, only respondent parallel-*present* targeted turns were observed, *and* that (2) when this behavior did not occur in the previous turns, only parallel-*absent* targeted turns were observed, or vice versa.

Our heuristic measures the degree of homogeneity using the entropy index. This index is measured *solely based on the homogeneity of the* ***retrieval behavior***. This index is calculated to equal 1 in conditions in which there is a lack of homogeneity, and it is equal to 0 when there is complete homogeneity. The entropy index for a single state is calculated according to the formula:

$$Entropy(S) = -\left(\frac{|S_{r=p}|}{|S|} * \log_2\left[\frac{|S_{r=p}|}{|S|}\right) - \left(\frac{|S_{r=a}|}{|S|} * \log_2\left[\frac{|S_{r=a}|}{|S|}\right)\right]\right), \qquad (1)$$

where S is the set of turns that we are interested in (e.g., the initial state of 3,488 turns), $S_{r=p}$ is the subset of these turns with the targeted retrieval behavior being present, and $S_{r=a}$ is the subset with the retrieval behavior being absent. Hence, and again using parallel retrievals as an example, the initial state of 1,744 parallel-present target turns and 1,744 parallel-absent target turns as a whole is calculated as having an $Entropy = 1$. On the other hand, a state with only parallel-present turns would have $Entropy = 0$.

As alluded to earlier, our behavior-select heuristic calculates the extent to which the presence and absence of each behavior affects the distribution for the respondent retrieval. This calculation measures the information gain *for each preceding occurring behavior.*

The information gain measures how applying the behavior, by splitting the turns into subsets where the behavior is present and absent, affects the entropy measured on the targeted final retrieval behavior. For each final retrieval behavior, two entropy indices are calculated, one index for the behavior-present state in which the behavior occurred in the previous three turns, and one index for the behavior-absent state:

$$Gain = Entropy(S) - \frac{|S_{b=p}|}{|S|} Entropy(S_{b=p}) - \frac{|S_{b=a}|}{|S|} Entropy(S_{b=a}), \qquad (2)$$

where S is the set of turns that we are interested in, $S_{b=p}$ is the subset of these turns with retrieval behavior-present, and $S_{b=a}$ is the subset with retrieval behavior-absent turns.

Based on the above equation, the preceding behavior with the highest information gain is the one that, when it is applied, provides the highest increase in homogeneity for the final retrieval behavior. Using the previous example, the highest information gain occurs when the presence of a preceding behavior always results in a final respondent parallel behavior being present, and the absence of a preceding behavior always results in the absence of a final respondent parallel behavior (or vice versa). Such a result would "zero out" terms 2 and 3 in the equations leading to $Gain = 1 - 0 - 0 = 1$.

Continue-Growing Heuristic

Once an initial behavior is identified that produces the largest information gain, we need to decide whether to continue growing the series adding additional behaviors. To address the homogenous-state stop case, our heuristic considers both (1) the branch in which a previously identified behavior was present in the preceding three turns, and (2) the branch in which a previously identified behavior was *not* present in the preceding three turns. The goal is to allow the decision-tree algorithm to proceed down any branch until the group achieves an $Entropy = 0$ when all the turns share a homogenous state. Any verbal behavior that is identified at any step as maximizing information gain is contingent on the presence or absence of behaviors identified during previous steps within that branch. As alluded to earlier, this allows the algorithm to produce a hierarchical network of series that each consist of steps with the isolation of specific behaviors whose presence or absence is required for each of the additional subsequent steps.

To address the small-group stop case, our heuristic uses a criterion rule, for each of the four analyses separately, which requires that a minimum of five percent (or 80) targeted turns need to contain a respondent retrieval for any specific behavior to be retained (either in the behavior-present state or behavior-absent state). Implementing the criterion rule resulted in a network of four steps and five different behavior series for parallel and timing retrievals (see Tables 1 and 2), and two steps and three series for both duration and sequential retrievals (see Tables 3 and 4).

Multiple-Behavior Series Results

For parallel retrievals (see Table 1), in the three turns (two interviewer and one respondent) that preceded the targeted respondent turns, four behaviors were found to discriminate between the occurrence and nonoccurrence of these retrieval strategies in the targeted turns: respondent timing in the second turn, and interviewer parallel, timing, and duration probes in the first or third turns. *Parallel probes* are defined as verbal behaviors in

Table 1. Behavior series and their statistics: parallel retrieval.

		Series				
Step	Behavior	I	II	III	IV	V
1	R Timing (2nd turn)	P	A	A	A	A
2	I Parallel		P	A	A	A
3	I Timing			P	A	A
4	I Duration				P	A

		Series				
	Statistics	I	II	III	IV	V
	N turns	340	202	594	124	484
	Proportion Turns	.195	.116	.341	.071	.278
	Series Ratio	.744	.795	.537	.582	.332

Notes: P = Behavior-Present; A = Behavior-Absent.

which interviewers use a contemporaneous event in another life domain as an anchor to cue respondents in the remembering of a domain-relevant event, *timing probes* asked when an event started or stopped and *duration probes* consist of interviewers asking how long an event occurred. Series I in Table 1 consists of at least one respondent timing retrieval in the second preceding turn, and this behavior is not at all constrained by the presence or absence of any other behaviors within the preceding three turns. Series II requires that there is no respondent timing retrieval behavior in the second turn, but that there is an interviewer parallel probe in either the first and third preceding turns. Series III is marked by the absence of respondent timing and interviewer parallel behaviors, and the presence of an interviewer timing probe. In Series IV, there must be an absence in the three preceding turns of respondent timing and interviewer parallel and timing behaviors, but the presence of an interviewer duration probe. Series V requires the absence of all four of these behaviors in the preceding three turns.

Table 2. Behavior series and their statistics: timing retrieval.

		Series				
Step	Behavior	I	II	III	IV	V
1	I Duration	P	P	A	A	A
2a	I Timing	A	P			
2b	R Timing (2nd turn)			P	A	A
3	I Data Elements				P	A

		Series				
	Statistics	I	II	III	IV	V
	N turns	541	160	287	350	1483
	Proportion Turns	.192	.057	.102	.124	.526
	Series Ratio	.780	.608	.651	.327	.467

Notes: P = Behavior-Present; A = Behavior-Absent.

Table 3. Behavior series and their statistics: duration retrieval.

		Series		
Step	Behavior	I	II	III
1	R Timing (2nd turn)	P	A	A
2	I Timing		P	A
		Series		
	Statistics	I	II	III
	N turns	200	442	549
	Proportion Turns	.168	.371	.461
	Series Ratio	.694	.562	.420

Notes: P = Behavior-Present; A = Behavior-Absent.

Table 1 also includes statistics that are associated with each series. The number of turns out of the total of 1,744 that include a respondent parallel retrieval are provided, as is the proportion (Series *N*/1744). As can be seen, Series III, which included the presence of interviewer timing, accounted for the most turns, and Series IV the least. In addition, the series ratio indicates the extent to which each series discriminated between turns with and without respondent parallel retrievals, with the higher values indicating greater discriminability. The ratio is calculated as the number of turns with respondent parallel retrieval in the targeted turn divided by the total number of turns that fit the series in the three turns that preceded the targeted turn. This ratio does not reflect the actual discriminability among all of the turns in the interview, as the ratio accounts only for the 1,744 nonrespondent retrieval turns that were randomly sampled.

Tables 2–4, which depict the data-mining results for respondent timing, duration, and sequential behaviors respectively, can be interpreted in the same way. For the most part these behaviors include interviewer or respondent timing and duration behaviors. Table 2 reveals in Series IV that interviewer data-element probes, in which interviewers ask for

Table 4. Behavior series and their statistics: sequential retrieval.

		Series		
Step	Behavior	I	II	III
1	R Timing (2nd turn)	P	A	A
2	R Parallel (2nd turn)		P	A
		Series		
	Statistics	I	II	III
	N turns	163	90	512
	Proportion Turns	.213	.118	.669
	Series Ratio	.751	.732	.310

Notes: P = Behavior-Present; A = Behavior-Absent.

detailed information such as the names of persons or employers, are predictive of respondent timing retrievals in the fourth turn.

All of the data-mining results (see Tables 1–4) consistently reveal that those behaviors that best predict the occurrence of respondent retrievals are either interviewer retrieval probes or other types of respondent retrievals. Importantly, retrieval behaviors are a subset of possible behavior types that were included in the analysis; they included interviewer and respondent conversational and rapport behaviors (e.g., clarifications, digressions), interviewer feedback behaviors that followed responses (e.g., "thank-you"), and respondent cognitive-difficulty behaviors (e.g., requests for question repeats). These data-mining results confirm the general finding from factor-analytic approaches that the same types of verbal behaviors tend to cluster together (Belli et al. 2001; Belli et al. 2013; Belli et al. 2004), but our data-mining results provide greater detail on the actual sequencing of behaviors as they occur in the exchanges between interviewers and respondents.

3.2. Validation Regression Models

The calendar retrospective reports were validated against these same respondents' reports, which had been provided in annual PSID interviews. Twelve cases suffered from processing errors, making the comparison of calendar retrospective reports to panel reports unfeasible, resulting in 153 validated cases. In this validation, we tested four domains in the reporting of residential, relationship, and labor histories; one was associated with residence, a second with marriage. With labor histories, we tested both employment and unemployment. These domains were selected as they each were designed to ask respondents to provide retrospective reports of objective facts. For each of these domains, we calculated respective measures of discrepancy; these were calculated as the proportion of years in which there was no match in status between the calendar and panel reports.

We tested logistic regression models to determine relationships between discrepancy and behaviors. In order to demonstrate that parallel, duration, timing, and sequential retrievals were associated with discrepancy, each of the respondent retrieval behaviors were tested. Models were examined for each of the domains separately and for each behavior separately, leading to sixteen analyses. In each model, per case, the discrepancy measure was regressed on the number of retrieval behaviors that had occurred up to and including the point at which each respective domain had been finalized during the interview, an experiential complexity measure based on the number of status changes for each respective domain in the panel data, a term for the interaction of the number of retrievals with experiential complexity, and control variables that included interviewer age, gender, and years of interviewing experience, and respondent age, gender, race, and years of education. The interaction term was included to determine whether the association of behavior with discrepancy hinged on those respondents who have more complicated histories; it may be the case that advantages of retrieval behaviors, if any, would only exist when retrieval is more difficult because the respondent has a complicated past (see Belli et al. 2013). If the model revealed a significant respondent retrieval behavior by experiential complexity interaction, a regions-of-significance (ROS) analysis was performed to determine at what level of experiential complexity the retrieval behavior was significantly associated with discrepancy. If the model did not reveal a significant interaction, the same core regression model without the interaction term was tested to determine whether the number of retrieval behaviors revealed a significant main effect on discrepancy.

To account for clustering of respondents within interviewers, covariance matrices were inflated using the estimated interviewer design effect for residence (*deff* = 1.64), marriage (*deff* = 2.00), and employment (*deff* = 1.45) discrepancy measures. Due to this clustering, it is appropriate to estimate the degrees of freedom used in significance tests as the number of interviewers – 1 = 12 (see Belli et al. 2013). The estimated design effect for the unemployment discrepancy measure was slightly less than 1 (0.97), indicating there was no interviewer clustering effect; hence, there was no need to inflate the covariance matrices.

Table 5 presents the results of the regression models, testing for interaction effects and their accompanying ROS results when significant. A greater number of respondent parallel retrieval behaviors is associated with less discrepancy in reports of being employed when respondents experienced greater experiential complexity. However, follow-up main-effects analyses revealed that a higher number of parallel retrievals is associated with greater discrepancy in reports of unemployment ($\beta = 0.014$, $SE = 0.006$, $p = .04$). As for timing retrieval behaviors, their greater propensity is associated with greater discrepancy in reports of residence and unemployment when respondents have less experiential complexity, but less discrepancy in reports of employment and unemployment when respondents have higher experiential complexity. Follow-up main-effects analyses also reveal that the greater number of timing retrievals is associated with less discrepancy in reports of marriage ($\beta = -0.061$, $SE = 0.015$, $p < .001$). With duration retrieval behaviors, their greater number is associated with less discrepancy in the reporting of employment when there is greater experiential complexity, and they also reveal less discrepancy in reports of marriage ($\beta = -0.091$, $SE = 0.032$, $p = .02$). Finally, a greater prevalence of sequential retrieval behaviors is associated with greater discrepancy in the reports of a) residence and unemployment with lower experiential complexity, and b) marriage with higher experiential complexity. They are also associated with less discrepancy in the reports of marriage with less experiential complexity and unemployment with higher experiential complexity. Overall, results indicate that respondents' engagement in timing and duration retrieval behaviors is beneficial to the accuracy of retrospective reports, especially when experiential complexity is high, but that engagement in parallel and sequential retrieval behaviors is mixed and nuanced in terms of data quality.

Having demonstrated the associations between each of the retrieval behaviors and discrepancy across the four domains, we tested logistic regression models to examine each of the series (see Tables 1–4) for each of the domains. These models included the same control variables and inflation of covariance matrices that were included in the models testing for respondent retrievals. In each model, discrepancy was regressed on the number of fourth-turn retrieval behaviors per case that met each interactional series as identified in Tables 1–4 (i.e., the five series for parallel, the five for timing, the three for duration, and the three for sequential retrievals). Only the retrieval behaviors that occurred up to and including the point in which each domain was being interviewed were included in the analysis.

Table 5. Logistic regression coefficients for the interaction of respondent retrieval behaviors and experiential complexity on discrepancy, and percentiles of experiential complexity associated with significantly less and greater discrepancy.

	Interaction Parameters			Percentiles of experiential complexity in which a greater number of respondent retrieval behaviors (in comparison to a fewer number) is associated with significantly	
Domain	Beta	SE	p	Less discrepancy	Greater discrepancy
			Parallel		
Residence	−0.090	0.061	ns		
Marriage	−0.345	0.576	ns		
Employment	−0.180	0.075	0.013	80.41–100	
Unemployment	−0.140	0.069	ns		
			Timing		
Residence	−0.151	0.053	0.047		0–51.57
Marriage	0.846	0.580	ns		
Employment	−0.178	0.065	0.006	30.22–100	
Unemployment	−0.241	0.047	<0.001	79.77–100	0–60.36
			Duration		
Residence	−0.177	0.134	ns		
Marriage	−1.827	1.177	ns		
Employment	−0.543	0.152	0.001	31.68–100	
Unemployment	−0.142	0.122	ns		
			Sequential		
Residence	−0.244	0.092	0.021		0–21.61
Marriage	5.920	1.359	<0.001	0–91.34	99.03–100
Employment	−0.066	0.112	ns		
Unemployment	−0.716	0.145	<0.001	72.08–100	0–52.81

Table 6a presents the interaction parameters of the regression models with parallel retrievals in the fourth turn and their accompanying ROS results when significant. There are no significant interaction effects for unemployment. Results demonstrate that for Series II in which an interviewer parallel probe occurs in the first or third turn, reports of employment are at greater discrepancy at higher levels of experiential complexity, while for Series III in which a respondent parallel is preceded by an interviewer timing probe, reports of employment are at greater discrepancy at lower levels of experiential complexity, but at less discrepancy at higher levels of complexity. For Series V, which is marked by a lack of preceding behaviors, reports of residence demonstrate greater discrepancy with lower complexity, and less discrepancy with higher complexity. There are also significant main effects: Series I, which includes a preceding respondent timing retrieval, Series IV, which has a preceding interviewer duration probe, and Series V are all

Table 6a. *Logistic regression coefficients for the interaction of interviewer-respondent sequences and experiential complexity on discrepancy, and percentiles of experiential complexity associated with significantly less and greater discrepancy: parallel.*

	Residence			Marriage			Employment			Unemployment		
		Discrepancy			Discrepancy			Discrepancy			Discrepancy	
Series	Beta (SE)	Less	Greater	Beta (SE)	Less	Greater	Beta (SE)	Less	Greater	Beta (SE)	Less	Greater
I	− 0.212 (0.291)			1.010 (2.286)			− 0.184 (0.230)			0.380 (0.256)		
II	− 0.133 (0.202)			1.484 (2.283)			0.543 (0.238)*		91.21–100	0.139 (0.227)		
III	− 0.166 (0.159)			− 0.714 (1.794)			− 0.852 (0.170)**	85.88–100	0–61.62	0.110 (0.158)		
IV	− 0.100 (0.212)			− 0.303 (1.716)			− 0.604 (0.507)			0.478 (0.226)		
V	− 1.189 (0.454)*	98.71–100	0–35.02	− 7.620 (5.466)			− 0.341 (0.303)			0.525 (0.388)		

$*p < .05$; $**p < .01$; $***p < .001$.

associated with less discrepancy in reports of marriage ($\beta = -0.138$, $SE = 0.059$, $p = .04$; $\beta = -0.120$, $SE = 0.047$, $p = .03$; $\beta = -0.395$, $SE = 0.124$, $p = .01$, respectively), and Series I is also associated with less discrepancy in reports of employment ($\beta = -0.091$, $SE = 0.028$, $p = .01$). Results demonstrate that the effectiveness of parallel retrieval behaviors is dependent on the preceding behavioral context. When preceded by a respondent timing behavior (Series I) or interviewer duration probe, data quality is improved; when preceded by an interviewer timing probe (Series III) or when there is no preceding behavior (Series V), data quality is improved when there is a more demanding retrieval task (higher experiential complexity); and when preceded by an interviewer parallel probe (Series II), data quality is worsened, especially with a more demanding retrieval task.

Results are also nuanced by which behaviors precede timing retrieval behaviors (see Table 6b). Series I, in which an interviewer duration probe precedes a timing retrieval behavior, reveals greater discrepancy at lower experiential complexity in reports of residence, but greater discrepancy at higher experiential complexity in reports of employment; in reports of employment, Series I also reveals less discrepancy with lower experiential complexity. Series II, in which a fourth-turn respondent timing behavior is preceded by both interviewer duration and timing probes, reveals less discrepancy with higher experiential complexity in reports of marriage, employment, and unemployment. With Series V, in which there are no preceding behaviors, there is less discrepancy with lower experiential complexity. There are also significant main effects: Series I and IV with marriage ($\beta = -0.157$, $SE = 0.037$, $p = .001$; $\beta = -0.093$, $SE = 0.029$, $p = .01$, respectively), and Series III, IV, and V with employment ($\beta = -0.208$, $SE = 0.043$, $p < .001$; $\beta = -0.065$, $SE = 0.012$, $p < .001$; $\beta = -0.221$, $SE = 0.049$, $p = .001$, respectively) all reveal less discrepancy. Taken together, results reveal that whereas the preceding occurrence of only duration probes (Series I) leads to mixed results with data quality, preceding duration probes in combination with timing probes (Series II) improve data quality when the retrieval task is difficult (higher experiential complexity). Moreover, preceding respondent timing behaviors (Series III) and preceding interviewer data-element probes (Series IV) improve data quality, whereas no preceding behaviors (Series V) leads to mixed results concerning data quality.

Duration retrieval behaviors at the fourth turn (see Table 6c) are noted for generally being associated with better data quality, especially at higher experiential complexity, regardless of the preceding behaviors. Both Series I, in which there is a preceding respondent timing behavior, and Series III, in which there are no preceding behaviors, are associated with less discrepancy at higher experiential complexity for employment reports; Series III also reveals less discrepancy at higher complexity and greater discrepancy at lower complexity with reports of marriage. As for Series II, in which there is a preceding interviewer timing probe, there are significant main effects in which there is less discrepancy with reports of both marriage ($\beta = -0.197$, $SE = 0.064$, $p = .01$) and employment ($\beta = -0.109$, $SE = 0.030$, $p < .01$). Table 6d reveals results for respondent sequential behaviors at the fourth turn. With Series II, in which there is a preceding respondent parallel behavior, whereas reports of residence reveal less discrepancy at higher complexity and greater discrepancy at lower complexity, the opposite pattern is seen with reports of marriage. There is also a main effect with Series I in which there is a

Table 6b. Logistic regression coefficients for the interaction of interviewer-respondent sequences and experiential complexity on discrepancy, and percentiles of experiential complexity associated with significantly less and greater discrepancy: timing.

	Residence			Marriage			Employment			Unemployment		
		Discrepancy			Discrepancy			Discrepancy			Discrepancy	
Series	Beta (SE)	Less	Greater	Beta (SE)	Less	Greater	Beta (SE)	Less	Greater	Beta (SE)	Less	Greater
I	− 0.407 (0.136)**		0–82.40	0.199 (1.588)			− 1.114 (0.220)***	0–47.13	83.13–100	− 0.274 (0.129)		
II	1.046 (0.509)			− 12.364 (5.226)*	91.11–100		− 1.605 (0.492)**	56.06–100		− 1.068 (0.349)**	82.93–100	
III	− 0.280 (0.379)			− 1.040 (2.928)			0.015 (0.394)			0.166 (0.258)		
IV	− 0.144 (0.138)			1.836 (1.088)			− 0.309 (0.450)			− 0.034 (0.076)		
V	− 0.263 (0.377)			8.292 (3.799)*	0–86.99		− 0.209 (0.118)			− 0.115 (0.219)		

$*p < .05$; $**p < .01$; $***p < .001$.

Table 6c. Logistic regression coefficients for the interaction of interviewer-respondent sequences and experiential complexity on discrepancy, and percentiles of experiential complexity associated with significantly less and greater discrepancy: duration.

	Residence			Marriage			Employment			Unemployment		
		Discrepancy			Discrepancy			Discrepancy			Discrepancy	
Series	Beta (SE)	Less	Greater	Beta (SE)	Less	Greater	Beta (SE)	Less	Greater	Beta (SE)	Less	Greater
I	0.193 (0.482)			2.157 (4.263)			−2.397 (0.697)**	40.01–100		0.037 (0.321)		
II	−0.307 (0.273)			1.278 (2.613)			−0.548 (0.332)			−0.119 (0.204)		
III	−0.451 (0.267)			−12.095 (2.531)***	90.97–100	0–74.03	−0.837 (0.254)**	34.39–100		−0.105 (0.195)		

*$p < .05$; **$p < .01$; ***$p < .001$.

Table 6d. Logistic regression coefficients for the interaction of interviewer-respondent sequences and experiential complexity on discrepancy, and percentiles of experiential complexity associated with significantly less and greater discrepancy: sequential.

	Residence			Marriage			Employment			Unemployment		
		Discrepancy			Discrepancy			Discrepancy			Discrepancy	
Series	Beta (SE)	Less	Greater	Beta (SE)	Less	Greater	Beta (SE)	Less	Greater	Beta (SE)	Less	Greater
I	− 0.743 (0.382)			− 2.669 (3.827)			− 0.172 (0.396)			0.236 (0.351)		
II	− 0.310 (0.120)*	98.52–100	0–26.90	12.095 (1.937)***	0–91.11	94.49–100	− 0.016 (0.130)			− 0.060 (0.126)		
III	− 0.204 (0.428)			2.683 (4.568)			− 1.021 (0.496)			0.364 (0.508)		

*$p < .05$; **$p < .01$; ***$p < .001$.

preceding respondent timing behavior such that there is less discrepancy in reports of marriage ($\beta = -0.284$, $SE = 0.111$, $p = .03$).

4. Implications for Interviewing Research and Practice

Results present a nuanced and complicated pattern of interactions among interviewer and respondent verbal behaviors in impacting the quality of retrospective reports in calendar interviews. Of the retrieval behaviors examined, both respondent and interviewer timing behaviors and respondent duration behaviors demonstrated the most consistent association with higher data quality, especially when the retrieval task was more difficult as measured by experiential complexity. In terms of respondent timing behaviors, overall prevalence was associated with better data quality; data quality was improved when preceded by interviewer timing and data-elements probes and respondent timing retrievals, and data quality was also improved when respondent timing behaviors preceded respondent parallel, timing, duration, and sequential retrieval behaviors. As for interviewer timing behaviors, their occurrence facilitated better data quality and preceding respondent parallel, timing, and duration behaviors. Respondent duration behaviors also revealed better data quality with overall prevalence with heightened retrieval difficulty, and data quality was improved when preceded by respondent and interviewer timing, or by the absence of behaviors.

Mixed results in terms of data quality were found for respondent and interviewer parallel behaviors, and respondent sequential and interviewer duration behaviors. The overall prevalence of both respondent parallel and sequential behaviors was not consistently associated with improved data quality, even when examining only situations in which the retrieval task was difficult, nor did the presence of these behaviors when preceded by other behaviors, or when preceding retrieval behaviors, produce consistent results in terms of data quality. The preceding presence of interviewer parallel probes led to poor data quality outcomes when followed by respondent parallel retrieval behaviors. Interviewer duration probes, when alone in preceding respondent timing retrieval behaviors, led to mixed results with data quality.

These results have implications for interviewer behaviors in calendar interviews. Whereas interviewer timing probes are to be encouraged, interviewer parallel probes are to be discouraged. As for interviewer duration probes, they appear to be effective only when used in combination with interviewer timing probes. Respondent timing strategies are also to be encouraged, and their occurrence is facilitated by interviewer timing, duration, and data-elements probes, although, as noted above, interviewer duration probes should not be administered alone. The encouragement of effective respondent duration strategies is also facilitated by interviewer timing probes.

One troubling aspect of providing advice in encouraging interviewer and respondent timing behaviors is that although heightened prevalence is associated with better data quality when the retrieval task is difficult, in some situations there is also poorer data quality when the retrieval task, as measured by lower levels of experiential complexity, is relatively easy. Such a pattern has also been observed by Belli et al. (2013), who speculate that some respondents experience general difficulty in remembering their pasts, and that interviewers are more prone to unsuccessfully use retrieval probes as an attempt to improve these respondents' memories. Accordingly, it may be advisable to attempt introducing screener questions to assess how much status changes have occurred in respondents' pasts, and to encourage interviewer and respondent timing behaviors for those respondents whose pasts are more complicated.

Another caveat in terms of attempts at implementing more successful interviewer training regimens is that some level of better data quality is due to respondents engaging in retrieval strategies on their own. For example, although respondent timing retrieval behaviors appear to be encouraged by interviewer timing, duration, and data-element probes, they also may occur spontaneously, and hence their benefits to data quality may be present only among a certain subset of respondents or circumstances in which interviewer probing has no impact. As other examples, both respondent parallel and duration behaviors, when preceded by no behaviors, are associated with better data quality especially with respondents who have complicated pasts, and hence, appear to be driven by respondents on their own.

5. Theoretical Considerations

A major surprise in our results is the lack of solid evidence that interviewer and respondent parallel behaviors improve the quality of retrospective reports. Much of the theoretical rationale of calendar interviewing has hinged on the notion that its implementation encourages the occurrence of effective cuing mechanisms, especially parallel behaviors (for examples see Balán et al. 1969; Belli 1998; Belli and Callegaro, 2009; Yoshihama et al. 2002). The finding that interviewer parallel probes lead to poorer retrospective data quality is particularly contrary to theoretical expectations. Gaining a better understanding of the role of parallel associations in human autobiographical memory may assist in determining how such associations impact accuracy.

Psychologists who have theorized about the structure of autobiographical knowledge have differed in opinion on whether parallel associations exist across contemporaneous events from different autobiographical domains or themes. On the one hand, theories of autobiographical memory that incorporate associations among different life domains are supported by the existence of respondent parallel cuing, as observed by Bilgen and Belli (2010) and as evident from the results reported in this article. Specifically, Barsalou (1988) observed that persons will follow parallel tracks of events that associate contemporaneous events across themes, such as events that encompass a project such as school being associated with contemporaneous events of being with one's family. He projected that such associations could exist between different events from different life domains. Similarly, Means, Loftus, and colleagues (Means and Loftus 1991; Means et al. 1989) observed that individuals would jump between work and health events when answering questions about their memory for health visits.

On the other hand, the presence of these parallel associations is not emphasized in some theories of autobiographical memory that highlight hierarchical associations among events within the same life domain (Conway and Bekerian 1987; Conway 1996). Specifically, Conway and Bekerian propose the existence of Autobiographical Memory Organization Packets (A-MOPs) that hierarchically organize more specific episodic events within abstract lifetime periods. As these A-MOPs are thematic with respect to encapsulating hierarchies consisting of different life domains, such as one's relationships versus one's career, an autobiographical memory structure consisting only of A-MOPS would not predict events belonging within one life domain to be cued by contemporaneous events that had occurred in a different life domain.

Overall, our results point to a compromise between these views. It may be the case that direct parallel associations are somewhat uncommon, and hence, that benefits from respondents' use of parallel retrievals may arise, but only when these direct parallel associations exist. However, directing respondents to engage in parallel retrieval through the use of parallel probes may divert respondents from more beneficial within-domain associations and increase task difficulty, leading to poorer autobiographical remembering.

6. Limitations and Other Considerations

Although results suggest that interviewer parallel probing does more harm than good and hence ought to be discouraged in interviewer training, the observational nature of this research means that these causal inferences are tentative. It may be possible, for example, that interviewers are more likely to engage in parallel probing with respondents who exhibit poor memory, and that the association of parallel probing with poorer data quality is the outcome of respondents who are not able to remember their pasts very well. It may also be the case that our data are limited in that they arose from telephone interviewing in which the calendar was not observed by respondents, and the use of other data-collection modes in which respondents can view the calendar may lead to overall benefits from parallel associations.

In addition, as our research is not experimental, more definitive answers regarding the impact of parallel probing could be gained through experimental work in which interviewers are either encouraged or discouraged to use parallel and other types of probes. Of course, the concern with making causal inferences also applies to timing probes, which were found to be associated with better quality data.

Results are also limited in that they have only examined the value of data-mining techniques with retrieval behaviors in calendar interviews, and with a relatively small sample. Extensions of data mining should also be applied to other behaviors in both calendar and conventional questionnaire interviewing instruments, especially those of a more direct motivational flavor. As for calendar interviews, the various series of behaviors that will lead to respondent rapport behaviors, such as digressions and laughter, may be of particular interest, as the use of rapport has been shown to be associated with better retrospective reporting in some domains but not others (Belli et al. 2013). As for conventional questionnaires, question-answer series that lead to behaviors signifying that respondents are having cognitive problems with the question may be especially informative, given that these behaviors have often been associated with poorer data quality (Belli and Lepkowski 1996; Draisma and Dijkstra 2004; Dijkstra and Ongena 2006; Dykema et al. 1997).

These are but examples, of course. The key message to take away is that data-mining techniques can be used in behavior-coding analyses to uncover those series of behaviors that produce key data-quality relevant behaviors. In combination with regression techniques, it can also be determined which of these series are associated with better or poorer data quality. The results from these investigations are important theoretically in the understanding of cognitive and communicative processes, and they have implications for interviewer training and in the development of best interviewing practices.

7. References

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

Barsalou, L.W. 1988. "The Content and Organization of Autobiographical Memories." In *Remembering Reconsidered: Ecological and Traditional Approaches to the Study of Memory*, edited by U. Niesser and E. Winograd, 193–243. New York: Cambridge University Press.

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. Doi: http://dx.doi.org/10.1080/741942610.

Belli, R.F. 2014. "Autobiographical Memory Dynamics in Survey Research." In *SAGE Handbook of Applied Memory*, edited by T.J. Perfect and D.S. Lindsay, 366–384. Los Angeles: Sage.

Belli, R.F., I. Bilgen, and T. Al Baghal. 2013. "Memory, Communication, and Data Quality in Calendar Interviews." *Public Opinion Quarterly* 77: 194–219. Doi: http://dx.doi.org/10.1093/poq/nfs099.

Belli, R.F. and M. Callegaro. 2009. "The Emergence of Calendar Interviewing: A Theoretical and Empirical Rationale." In *Calendar and Time Diary Methods in Life Course Research*, edited by R.F. Belli, F.P. Stafford, and D.F. Alwin, 31–52. Thousand Oaks, CA: Sage.

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

Belli, R.F. and J.M. Lepkowski. 1996. "Behavior of Survey Actors and the Accuracy of Response." In *Proceedings of the Conference on Health Survey Research Methods*, June, 1995, Breckenridge, CO, 69–74. DHHS Publication No. (PHS) 96-1013.

Belli, R.F., J.M. Lepkowski, and M.U. Kabeto. 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*, edited by M.L. Cynamon and R.A. Kulka, 197–203. DHHS Publication No. (PHS) 01-1013. Hyattsville, MD: U.S. Government Printing Office.

Belli, R.F., L. Smith, P. Andreski, and S. Agrawal. 2007. "Methodological Comparisons between CATI Event History Calendar and Conventional Questionnaire Instruments." *Public Opinion Quarterly* 71: 603–622. Doi: http://dx.doi.org/10.1093/poq/nfm045.

Bilgen, I. and R.F. Belli. 2010. "Comparison of Verbal Behaviors between Calendar and Standardized Conventional Questionnaires." *Journal of Official Statistics* 26: 481–505.

Brenner, M. 1982. "Response-Effects of 'Role Restricted' Characteristics of the Interviewer." In *Response Behavior in the Survey Interview*, edited by W. Dijkstra and J. van der Zouwen, 131–165. London: Academic Press.

Conway, M.A. 1996. “Autobiographical Knowledge and Autobiographical Memories.” In *Remembering Our Past: Studies in Autobiographical Memory*, edited by D.C. Rubin, 67–93. New York: Cambridge University Press.

Conway, M.A. and D.A. Bekerian. 1987. “Organization in Autobiographical Memory.” *Memory and Cognition* 15: 119–132. Doi: http://dx.doi.org/10.3758/BF03197023.

Draisma, S. and W. Dijkstra. 2004. “Response Latency and (Para)Linguistic Expressions as Indicators of Response Error.” In *Methods for Testing and Evaluation of Survey Questionnaires*, edited by S. Presser, J.M. Rothgeb, M.P. Couper, J.T. Lessler, E. Martin, J. Martin, and E. Singer, 131–147. Hoboken, NJ: Wiley.

Dijkstra, W. and W. Ongena. 2006. “Question-Answer Sequences in Survey Interviews.” *Quality and Quantity* 40: 983–1011. Doi: http://dx.doi.org/10.1007/s11135-005-5076-4.

Dykema, J., J.M. Lepkowski, and S. Blixt. 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*, edited by L. Lyberg, P. Biemer, M. Collins, E. de Leeuw, C. Dippo, N. Schwarz, and D. Trewin, 287–310. New York: J.W. Wiley and Sons.

Freedman, D., A. Thornton, D. Camburn, D. Alwin, and L. Young-DeMarco. 1988. “The Life History Calendar: A Technique for Collecting Retrospective Data.” In *Vol. 18 of Sociological Methodology*, edited by C.C. Clogg, 37–68. San Francisco: Jossey-Bass.

Glasner, T. and W. van der Vaart. 2009. “Applications of Calendar Instruments in Social Surveys: A Review.” *Quality and Quantity* 43: 333–349. Doi: http://dx.doi.org/10.1007/s11135-007-9129-8.

He, H. and E. Garcia. 2009. “Learning from Imbalanced Data.” *IEEE Transactions on Knowledge and Data Engineering* 21: 1263–1284. Doi: http://dx.doi.org/10.1109/TKDE.2008.239.

Means, B. and E.F. Loftus. 1991. “When Personal History Repeats Itself: Decomposing Memories for Recurring Events.” *Applied Cognitive Psychology* 5: 297–318. Doi: http://dx.doi.org/10.1002/acp.2350050402.

Means, B., A. Nigam, M. Zarrow, E.F. Loftus, and M.W. Donaldson. 1989. “Autobiographical Memory for Health-Related Events.” *Vital and Health Statistics*. DHHS Publication No. PHS 89-1077, Series 6, Number 2. Washington, DC: US Government Printing Office.

Weiss, G. and F. Provost. 2001. “The Effect of Class Distribution on Classifier Learning: An Empirical Study.” Rutgers University Technical Report ML-TR-44.

Witten, I., E. Frank, and M. Hall. 2011. *Data Mining: Practical Machine Learning Tools and Techniques*. Amsterdam: Elsevier.

Yoshihama, M., K. Clum, A. Crampton, and B. Gillespie. 2002. “Measuring the Lifetime Experience of Domestic Violence: Application of the Life History Calendar Method.” *Violence and Victims* 17: 297–317. Doi: http://dx.doi.org/10.1891/vivi.17.3.297.33663.

Received April 2015
Revised December 2015
Accepted May 2016