Combining Paradata and Survey Responses to Identify Sources of Measurement Error in Medical Event Reporting

Andrew Mercer[1], Weijia Ren[1], Steven Machlin[2]
Frederick Rohde[2], Jeffrey Gonzalez[2]
[1]Westat, 1600 Research Boulevard, Rockville, MD 20850
[2]Agency for Healthcare Research and Quality, 540 Gaither Road, Rockville, MD 20850

Abstract

The ability to observe within-subject change over time is the primary objective of most panel surveys. When characteristics of the data collection process systematically affect reporting differently at different times, it becomes difficult to differentiate true change from measurement error. The Medical Expenditure Panel Survey (MEPS) employs an overlapping panel design in which new cohorts enter the survey every January and are interviewed five times covering a cumulative two-year reference period. Underreporting is a perennial concern for household surveys and this concern may be exacerbated in panel surveys because of issues such as panel conditioning (Kalton et al 1989). In particular, a review of the literature pertaining to the accuracy of household-reported healthcare utilization data suggests that medical events tend to be underreported (Bhandari and Wagner 2006; Zuvekas and Olin 2009). Separate MEPS panels consistently exhibit a pattern of disproportionately high medical event reporting in the first round relative to all subsequent rounds and an additional decline at the final round of data collection. The fact that this pattern persists across separate panels suggests that these differences may reflect measurement error. Steps to repair this error will depend on its cause. One hypothesis is that respondents reduce their reporting in Round 2 in order to reduce burden. Alternatively, the error may be cognitive in origin, with longer reference periods in Round 2 resulting in a greater level of forgetting on the part of the respondent. In this paper we compare the plausibility of these hypotheses for explaining changes in response patterns using both paradata and survey responses. We find no support for the hypothesis that burden leads to lower reporting, however, we do find a negative association between the length of the reference period and the level of reporting.

Key Words: Paradata, measurement error, panel surveys, panel conditioning, recall error

1. Introduction[1]

The Medical Expenditure Panel Survey Household Component (MEPS) is an ongoing, nationally representative face-to-face survey of households in the United States. The survey tracks Americans' health conditions, use of medical services and medical

[1] The views expressed in this article are those of the authors, and no official endorsement by the Department of Health and Human Services or the Agency for Healthcare Research and Quality is intended or should be inferred.

expenditures, as well as economic and demographic indicators. MEPS is conducted by Westat on behalf of the Agency for Healthcare Research and Quality and the National Center for Health Statistics (Westat 2012). MEPS uses an overlapping panel design, in which each year a new panel is selected from a frame consisting of households that completed the prior year's National Health Interview Survey (NHIS). Respondents are interviewed a total of five times approximately six months apart.

One of the main purposes of the MEPS is to measure the American public's medical utilization, or the number of times individuals interact with medical service providers of all types. As a longitudinal survey, MEPS is designed to measure medical utilization in the aggregate as well as how individuals' consumption of medical services changes over time. Previous research has raised concerns over the accuracy of self-reported use data. Specifically, the literature suggests that respondents imperfectly recall medical events and, on average, tend to underreport such events (Zuvekas 2011). MEPS has exhibited a pattern where the reported rate of medical utilization is systematically higher in the first round of data collection relative to all subsequent rounds with another decline observed in the final round. Panel conditioning, when a respondent's experience in early waves of data collection influences their reporting behaviors in later waves, is one possible explanation for this phenomenon. A second possibility is the fact that, with the exception of Round 5, Round 1 generally entails a shorter reference period than subsequent rounds of data collection, and that this easier recall task may result in respondents forgetting fewer events. This analysis evaluates rates of medical utilization reporting in the first two rounds of three separate MEPS panels in order to evaluate these two potential explanations.

Medical events refer to any kind of interaction between a person and some kind of medical service provider. They can be visits to a doctor's office, hospital (inpatient, outpatient and emergency room) or dentist, or interactions with another medical service provider such as a home health care worker or nurse practitioner[2]. Although they are recorded as medical events in MEPS, we exclude purchases of prescription medications in this analysis because the focus is on reporting of interactions with medical service providers.

In MEPS, a single respondent typically reports on behalf of the entire household. Respondents are not asked for an overall number of events, but rather are asked to identify and describe each medical event that occurred for each family member during the reference period. In Round 1, the reference period begins on January 1 and ends on the day of the first interview. In Rounds 2 through 4 the reference period covers the period of time between interviews, while in Round 5 the reference period extends from the prior interview through December 31 of the respondent's second year of participation in the survey. Because of the difficulty of reporting this type of information, respondents are encouraged to make use of medical records (e.g., explanation of benefits received from health insurance plans) to aid in enumerating medical events for themselves and other household members. In this paper, the medical utilization rate for a household represents an annualized frequency of medical events per person in the household and is formally defined as:

[2] In contrast to a per visit basis as for the other event types, home health utilization is reported on a monthly basis and other medical equipment and supplies (e.g., orthopedic items, hearing devices) are generally reported on an annual basis.

$$\text{Utilization Rate} = \frac{\text{\# Total Events - \# Prescription Med Events}}{\text{\# Persons In HH} \times \text{\# Days in Reference Period}}$$

Figure 1 depicts the mean utilization rate among households that completed all five rounds of data collection for MEPS panels 13, 14 and 15, which began in 2008, 2009 and 2010 respectively. The estimates are weighted to account for the complex design and nonresponse in Round 1. It is immediately clear that the utilization rate reported in Round 1 is consistently higher than in other rounds. Combining across panels, a design-adjusted paired t-test of the difference in means between Rounds 1 and 2 is highly significant ($p < 0.001$). Moreover, it does not appear to be seasonal or we would expect to see a comparable increase at Round 3, which takes place during the same time of year as Round 1.

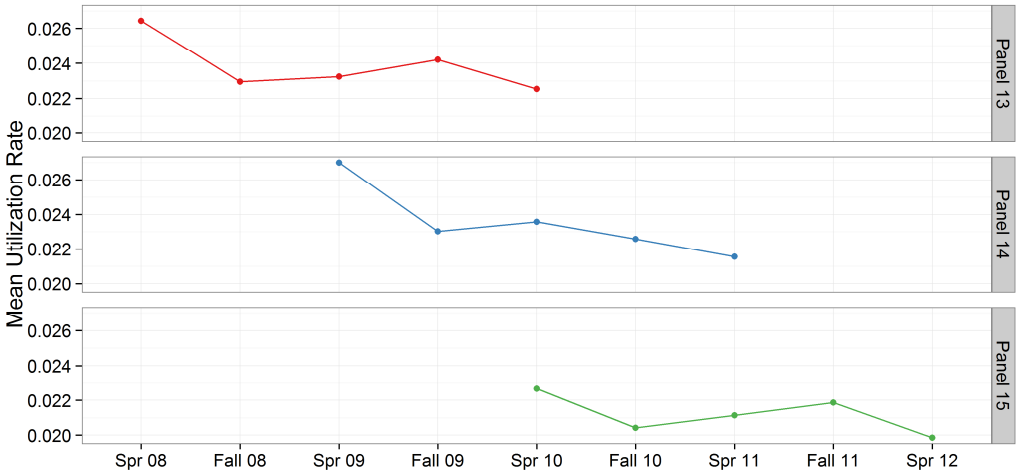

Figure 1: MEPS mean medical utilization rate by panel and round among households that completed all five rounds. Estimates are weighted to adjust for probability of selection and round 1 nonresponse.

Panel conditioning is a common worry for researchers involved in longitudinal studies. If the very act of participating in one round of a study can affect the results in subsequent rounds, then the study's results may not generalize to a larger target population as they are to some extent conditional upon participation in the study. Sometimes panel conditioning consists of changes in respondent behavior or attitudes produced by some survey stimulus (e.g. if asking about a topic causes the respondent to form new opinions about that topic). Survey burden has the potential to result in measurement error if it diminishes respondent motivation (Bradburn 1978). In this analysis we are concerned with the form of panel conditioning in which a burdensome experience in the first round of data collection may lead to lower motivation in subsequent rounds (See Cantor 2007 for a review of studies on panel conditioning). In MEPS, a large number of medical events in a household can potentially result in a particularly long interview. Each reported event generally involves complex follow up questions on sources of payment, payment amounts and charges, and requests for permission to obtain administrative records from the medical service provider. If some respondents found the first round to be particularly burdensome, it is possible that they will engage in satisficing in later rounds (Krosnick

1991). In the context of MEPS, satisficing behavior could mean that respondents fail to fully report all of their medical events in order to reduce the level of effort required to complete the interview.

Although we do not know how individual respondents perceived the level of burden in their Round 1 interview, we hypothesize that it should be correlated with the amount of time required to complete the interview. Several studies have looked at interview length as a measure of burden, although these have mostly pertained to attrition (Branden et al 1995, Zabel 1998, Hill and Willis 2001, Watson and Wooden 2009). The findings of these studies have been mixed, although Watson and Wooden (2009) found that attrition was highest for particularly short and particularly long surveys. We hypothesize that if burden is the explanation for the change in reporting, we should expect to see a pattern where over some threshold, longer interviews in Round 1 become associated with lower reporting in Round 2.

A second possible explanation for lower utilization reporting at Round 2 is that Round 1 simply presents a different and less difficult recall task than Round 2. One survey design feature that differs between Round 1 and Round 2 is the reference period. In Round 1, interviewers seek to contact newly sampled households quickly after an initial recruitment letter is sent in order to move forward with their initial workload. As a result, Round 1 reference periods tend to be shorter than Round 2 reference periods. Figure 2 depicts the distribution of reference periods for Rounds 1 and 2 for Panels 13, 14 and 15 combined. For Round 1 the average reference period is 93 days, whereas for Round 2 the mean is 167 days. Furthermore, while the distribution is in Round 2 is roughly symmetrical, the distribution for Round 1 is strongly skewed to the right indicating that most interviews have reference periods shorter than the mean.

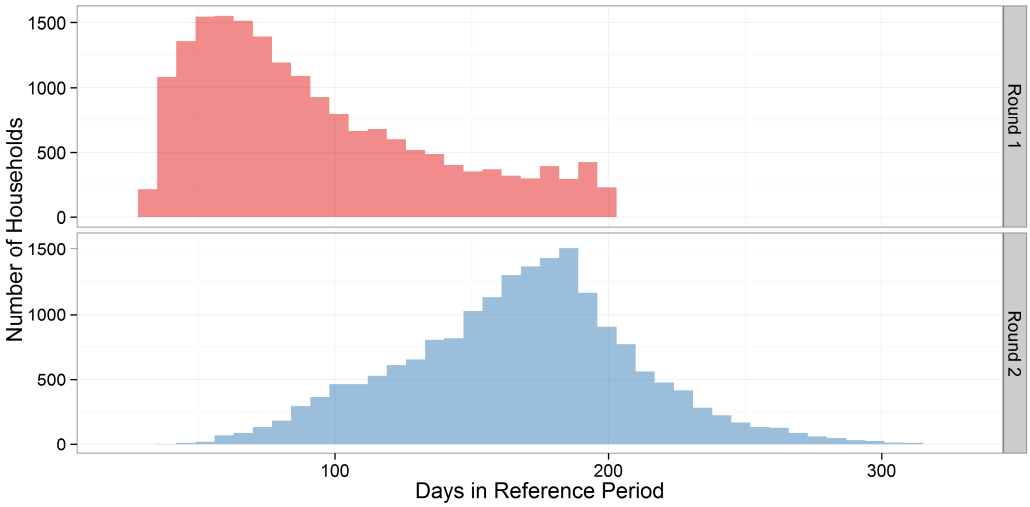

Figure 2: Distribution of respondent reference periods for MEPS Rounds 1 and 2

Researchers have found that when presented with a longer reference period, respondents face more difficulty in remembering individual events and are more likely to resort to forms of estimation and approximation when presented with questions about the frequency of certain behaviors. In particular, they are less likely to enumerate specific events (Blair and Burton 1987, Bradburn et al 1987). Warner et al (2005) found that a

three month long reference period results in more forgetting of medical events than a three to six week reference period on the NHIS, particularly for events that are minor and less memorable. Under this hypothesis, medical utilization reports are lower in Round 2, in part, because the longer reference period may make it more difficult for respondents to remember individual events and as a consequence they may be more prone to resort to forms of estimation when answering questions about medical utilization.

To evaluate these hypotheses, we pose two specific research questions which focus only on the changes in medical utilization reporting between Rounds 1 and 2.

1. Is lower medical utilization reporting in Round 2 associated with longer interview durations in Round 1?

2. Is lower medical utilization reporting in Round 2 associated with longer reference periods in Round 2?

2. Data and Methods

For this analysis, we use data on household characteristics and medical event reporting as well as paradata pertaining to 18,693 households that completed both Rounds 1 and 2 of MEPS Panels 13, 14 or 15. A small number of households missing data for key analytic variables were excluded. Households that changed size between Rounds 1 and 2 were also excluded in order to eliminate changes in utilization reporting caused by changes in household composition such as the passing away of a sick family member or the birth of a child.

We addressed both research questions using one linear regression model. To account for MEPS' complex sample design, we use design adjusted regression that incorporates the sample's stratification, clustering and weighting, and uses Taylor series linearization for variance estimation (Heeringa et al 2010). Because we are using a household-level file containing household summary data and paradata, the weights differ from those found in the publicly released person-level file. Here, the weights used are at the dwelling unit level and are adjusted for nonresponse that occurs in Round 1. They do not include additional adjustments or for nonresponse in later rounds. It is possible for multiple households to reside in the same dwelling – for example unrelated roommates are treated as separate reporting units in MEPS even though they reside in the same dwelling. In these instances, all households in the dwelling share the same weight.

The dependent variable is the utilization rate at Round 2, measured in events per day per person, denoted as R2Util in the equation below. The measure is then scaled up to 100 days to improve the readability of regression coefficients by reducing the number of leading zeroes in the output. To test the first hypothesis that longer interviews in Round 1 are associated with lower reporting in Round 2, we include a categorical variable indicating whether a household's Round 1 interview took up to 60 minutes, 61 to 90 minutes (R1Int61), 91 to 120 minutes (R1Int91), 121 minutes to 180 minutes (R1Int121), or over 180 minutes (R1Int181),. A categorical approach to the length of interview was chosen because the variable is highly skewed and to allow for nonlinearities in its relationship with utilization reporting. These particular categories were chosen because they create cells that are both large enough for analysis while remaining substantively meaningful. To test the hypothesis that longer reference periods are associated with lower

reporting in Round 2 relative to Round 1, we include the number of days in the Round 2 reference period as a covariate (DaysR2). Because the distribution of the Round 2 reference period is roughly symmetrical and does not show evidence of a nonlinear relationship with utilization reporting, we do not categorize or transform the variable.

$$\text{R2Util}_i = \beta_0 + \beta_1 \text{DaysR2}_i + \beta_2 \text{R1Int61}_i + \beta_3 \text{R1Int91}_i + \beta_4 \text{R1Int121}_i + \beta_5 \text{R1Int181}_i + \text{Control Variables} + \varepsilon_i$$

We also included several control variables in the regression model. These may each be associated with Round 2 utilization, days in the Round 2 reference period, and length of Round 1 interview. Specifically, we expect that those respondents with a high utilization rate in Round 1 are also more likely to have a high rate in Round 2, so we include the Round 1 utilization rate in the regression as a control. In order to control for differences between types of households, we also include other Round 1 variables associated with medical event reporting including the size of the household, the proportion of household members with chronic medical conditions, the proportion of household members aged 0 to 15, 16 to 64 and 65 or older, and if all, some or none of the household members have health insurance coverage. Households are assigned to the first race/ethnicity category for which to which any household member belongs in the following order: Asian non-Hispanic, Hispanic, black non-Hispanic or white/other non-Hispanic. Also included are characteristics of the interview process that are associated with the level of utilization reporting. Specifically, these are indicators for whether or not records were used to aid in recall in Rounds 1 or 2, whether or not the household required refusal conversion in Rounds 1 or 2 and the age of the respondent who reported on behalf of the household in Round 2.

3. Results

The results of the regression analysis are presented in Table 1. The hypothesis that burden in Round 1 produces lower reporting in Round 2 does not appear to be supported by this analysis. Because the coefficients for each time category increase monotonically, this model does not predict any increment where a longer interview in Round 1 would be expected to result in lower reporting in Round 2.

Table 1: Design Adjusted Regression Model Predicting Medical Utilization Rate in Round 2 of MEPS Data Collection

	Estimated regression coefficients			
Parameter	*Estimate*	*Std. Err.*	*T-value*	*p-value*
Intercept	1.195	0.146	8.18	<.0001
Days in R2 Ref. Prd.	-0.004	0.001	-5.99	<.0001
R1 Interview 61-90 Min	0.085	0.062	1.37	0.1722
R1 Interview 91-120 Min	0.324	0.105	3.09	0.0022
R1 Interview 121-180 Min	0.369	0.136	2.72	0.0072
R1 Interview Over 180 Min	0.380	0.118	3.21	0.0015
# Persons in HH	-0.198	0.022	-9.14	<.0001
% HH with Chronic Condition	0.497	0.071	6.98	<.0001
Asian HH	-0.211	0.067	-3.14	0.0019
Hispanic HH	-0.149	0.047	-3.17	0.0018
Black HH	-0.203	0.071	-2.85	0.0049
% HH Aged 0-15	0.206	0.102	2.02	0.0447

% HH Aged Over 64	0.643	0.217	2.96	0.0034
All HH Insured	0.278	0.045	6.14	<.0001
No Persons Insured	-0.265	0.062	-4.27	<.0001
Used Records R1	-0.413	0.072	-5.71	<.0001
Used Records R2	0.581	0.072	8.03	<.0001
Records R1 x Records R2	0.491	0.104	4.7	<.0001
Refusal Conversion R1	-0.005	0.087	-0.06	0.9536
Refusal Conversion R2	-0.410	0.123	-3.35	0.001
Respondent 18-44 yrs	-0.155	0.055	-2.8	0.0055
Respondent Over 64 Yrs	-0.319	0.180	-1.77	0.0788
Round 1 Utilization Rate x 100	0.431	0.027	15.97	<.0001
Panel 14	-0.030	0.060	-0.49	0.6212
Panel 15	-0.037	0.058	-0.64	0.5229
R-Square=0.36				
n=18,693, Strata=165, Clusters=371, DF=206				

For the second hypothesis, the length of the Round 2 reference period exhibits a statistically significant negative association with the Round 2 utilization rate. In particular, for every day added to the Round 2 reference period, we predict a 0.004 decrease (on average) in the Round 2 utilization rate. This effect is based on reference period and utilization data for Round 2, but the distribution of reference period length is different for Round 1. Therefore, to better understand the effect of reference period on utilization reporting we estimated what reporting would be if the length of the reference period for Round 2 was the same as for Round 1. To this end, we calculate a predicted value for each household where all independent variables remain unchanged in value except for the length of the Round 2 reference period. For each household, the value for the Round 2 reference period is set equal to that household's value for Round 1. The weighted mean of these predicted values represents the level of utilization reporting expected under this model if the distribution of reference periods among respondents in Round 2 was identical to that in Round 1. This general approach can be used as a heuristic for interpreting the substantive meaning of otherwise difficult to interpret regression coefficients (See Gelman and Hill 2007: pp. 101-104 for a similar example applied to logistic regression).

The results of this calculation are depicted in Figure 3 which compares the actual observed change in mean utilization rate between Round 1 and 2 in red with this hypothetical value in blue. It shows that under this model, the magnitude of the decline between rounds is reduced by approximately 90% if reference periods are held to their Round 1 value. Although this estimate is intended for aiding in interpretation and does not involve a formal test of statistical significance, it does suggest that differences in the length of the reference period could account for a substantial portion of the difference in utilization reporting between Rounds 1 and 2.

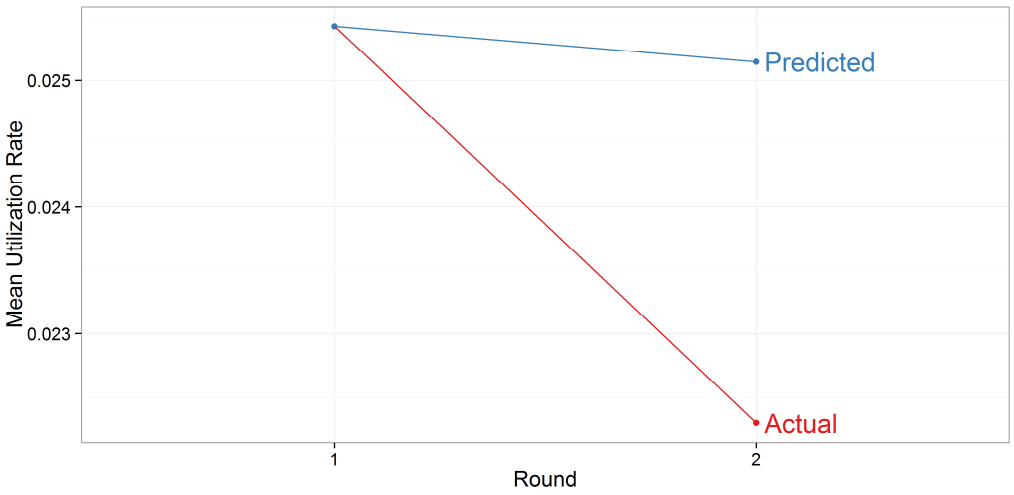

Figure 3: Comparison of actual change vs. change predicted by the regression model if R2 reference period lengths were equal to R1

4. Discussion

This analysis sought to evaluate two explanations for differential medical utilization reporting between Rounds 1 and 2 of MEPS. First, we wished to know if survey burden in Round 1 could be producing lower levels of utilization reporting in Round 2. The results of the regression analysis fail to provide evidence that longer Round 1 interviews are associated with lower medical utilization reporting in Round 2. Here we used survey length as a proxy for respondent burden. It should be noted however that survey length is only one dimension of respondent burden. Bradburn (1978) outlines several other dimensions of burden including respondent stress, required effort and the frequency of being interviewed, none of which are considered in this analysis. In the future, it would be worthwhile to obtain measures of these other dimensions of burden in MEPS in order to assess their potential impact on respondent reporting as well as other potential error sources such as respondent attrition.

Second, is a longer Round 2 reference period associated with lower reported utilization in Round 2? We have evidence from the regression analysis that a longer reference period in Round 2 is associated with lower Round 2 utilization reporting. Although this analysis does not definitively show a causal relationship, it does suggest that further investigation into the effects of reference period length is warranted. It is important to consider that although shorter reference periods may improve recall, as evidenced by Warner (2005), this may not be a panacea. Shorter reference periods could improve respondent recall, but would require more frequent interviews. In addition to increasing data collection costs, more frequent interviewing, one of Bradburn's other dimensions of burden, could lead to increased survey burden, greater attrition and may introduce new sources of measurement error. For example, Cohen and Burt (1985) found that more frequent interviews in the National Medical Care Expenditure Care Survey actually resulted in lower utilization reporting over the course of a year.

Although the net effect of all reporting errors in each round of data collection in MEPS is likely underreporting, it is possible that some forward telescoping occurs. Forward

telescoping is when respondents remember events as having occurred more recently than they did. This could potentially upwardly bias the number of events reported in Round 1. As we stated previously, future research should also include broadening the analysis to include all five rounds of MEPS data collection with a particular focus on understanding the additional decline in utilization reporting observed in the Round 5 interview. This analysis suggests that equalizing reference periods in Rounds 1 and 2 could lead to more similar levels of utilization reporting. If the effect of the reference period is consistent for all five rounds, it could be the case that greater control over the spacing of respondent interviews could lead to more consistent results without actually increasing the number of interviews. This analysis also looked at the overall number of reported events and only controls for different respondent characteristics. Future analysis should assess the extent to which the effects of reference period vary for different types of medical events or for different types of respondents. It is conceivable that some categories of respondent might benefit from a shorter reference period while others do not. Ideally this research would also include a randomized field experiment in order to better assess causal effects.

References

Bhandari, A. and T. Wagner. 2006. Self-reported utilization of health care services: Improving measurement and accuracy. In *Medical Care Research and Review.* 63, 217-235.

Blair, E. and S. Burton. 1987. Cognitive processes used by survey respondents to answer behavioral frequency questions. In *Journal of Consumer Research*. 14, 280-288.

Bradburn, N. 1978. Respondent burden. In *Proceedings of the American Statistical Association*. 35-40.

Bradburn, N.M., L.J. Rips, and S.K. Shevell. 1987. Answering autobiographical questions: The impact of memory and inference on surveys. In *Science*. 236, 157-161.

Branden, L., R.M. Gritz, and M. Pergamit. 1995. The effect of interview length on attrition in the National Longitudinal Study of Youth. In *National Longitudinal surveys Discussion Paper, No. 28*. Washington, DC: Bureau of Labor Statistics, U.S. Department of Labor. Retrieved September 10, 2013. http://bls.gov/osmr/pdf/nl950030.pdf

Cantor, D. 2007. A review and summary of studies on panel conditioning. In *Handbook of Longitudinal Research: Design Measurement and Analysis*, ed. S. Menard, Burlington, VT: Academic Press. 123-138.

Cohen, S.B. and V.L. Burt. 1985. Data collection frequency effect in the National Medical Care Expenditure Survey. In *Journal of Economic and Social Measurement*. 13, 125-51.

Gelman, A. and J. Hill. 2007. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge, MA: Cambridge University Press.

Heeringa, S., B.T. West, and P.A. Berglund. 2010. *Applied Survey Data Analysis*. Boca Raton, FL: Chapman & Hall/CRC.

Hill, D.H. and R.J. Willis. 2001. Reducing panel attrition: A search for effective policy instruments. In *The Journal of Human Resources*. 36, 416-438.

Kalton, G., D. Kasprzyk, and D.B. McMillen. 1989. Non-sampling errors in panel surveys. In *Panel Surveys*, eds. G. Kalton, D. Kasprzyk and D.B. McMillen. New York: John Wiley and Sons. 249-270.

Krosnick, J.A. 1991. Response strategies for coping with the cognitive demands of attitude measures in surveys. In *Applied Cognitive Psychology*. 5, 213-236.

Warner, M., N. Schenker, M.A. Heinen, and L.A. Fingerhut. 2005. The effects of recall on reporting injury and poisoning episodes in the National Health Interview Survey. In *Injury Prevention*. 11, 282-287.

Watson, N. and M. Wooden. 2009. Identifying factors affecting longitudinal survey response. In *Methodology of Longitudinal Surveys*, ed. P. Lynn. Chichester, UK: John Wiley and Sons. 157-181.

Westat. 2012. MEPS annual methodology report. Agency for Healthcare Research and Quality, Rockville, MD. Retrieved April 30, 2013. http://meps.ahrq.gov/mepsweb/data_files/publications/annual_contractor_report/hc_ann_cntrct_methrpt.shtml

Zabel, J.E. 1998. An analysis of attrition in the panel study of income dynamics and the survey of income and program participation with an application to a model of labor market behavior. In *The Journal of Human Resources*. 33, 479-506.

Zuvekas, S.H. 2011. The effects of recall length and reporting aids on household reporting of health care events in the Medical Expenditure Panel Survey. In *Journal of Economic and Social Measurement*. 36, 321-343.

Zuvekas, S.H. and G.L. Olin. 2009. Validating household reports of health care use in the Medical Expenditure Panel Survey. In *Health Services Research*. 44, 1679–1700.