Errors in the Recorded Number of Call Attempts and Their Effect on Nonresponse Adjustments Using Callback Models

Paul P. Biemer, Patrick Chen, and Kevin Wang

RTI International

1. Introduction

Nonresponse weighting adjustments use variables that are known for both respondents and nonrespondents. These data should be related to survey items of interest, as well as response propensity —i.e., the likelihood that a sample member will participate in the survey. One variable that is related to the latter is the numbers of attempts or callbacks interviewers make in order to obtain an interview, sometimes referred to as *level of effort* (LOE) data. LOE data may be call attempts on a phone survey or visits to a sampled unit in a field survey. Most survey organizations routinely store information on LOE while conducting surveys as part of the data collection management process.

As we show in this paper, LOE data are useful for nonresponse modeling provided they are reasonably accurate. Callback response models assume that all callbacks to potential respondents are recorded and can be defined and captured in a standard way. The estimated response propensity for an individual case is a function of the number of callbacks in that as the number of callbacks increases, the response propensity also increases. If the LOE data are not recorded accurately, estimates of the response propensity from callback models will be biased. In turn, survey estimates incorporating these response propensity adjustments will also biased.

This paper considers the accuracy of callback data and the effect of errors in these data on the parameter estimates obtained from callback modeling. Section 2 describes the callback model and establishes the notation that will be used in the paper. Section 3 summarizes the results of a study that applied a number of callback models to adjust for nonresponse in the National Survey of Drug Use and Health (NSDUH), a national face-to-face survey. Section 4 provides some findings from a companion study that evaluated the accuracy of the callback data from the NSDUH. Section 5 provides some preliminary results of a simulation study to evaluate the effects on the callback model parameters of errors in the number of call attempts. Finally, Section 6 summarizes our conclusions.

2. Model

The callback model considered in this work was proposed by Biemer, Chen, and Wang (2010) for the 2006 NSDUH. The reader is referred to that paper for the details of the analysis. Here we provide a brief summary of their approach and results.

Assume that following a call attempt, a case can be classified into three categories: (1) interviewed; (2) not interviewed, but contacted; and (3) no contact. Let n_{adg} denote the number of persons who were contacted for the first time at attempt a, and have final call disposition d, and who belong to group g, where $a = 1,\ldots,K$; $d = 1,2$; and $g = 1,\ldots,L$. Note that $\sum_a \sum_d \sum_g n_{adg} = n$, the overall sample size. Let n_{ad+} denote summation across groups (i.e., $n_{ad+} = \sum_g n_{adg}$). For example, n_{K3+} is the number of cases that were still not contacted after K call attempts. Finally, assume that cases are not terminated prematurely (e.g., censored); i.e., all cases having final "not

contacted" disposition were attempted *K* times. The model in Biemer, Chen, and Wang (2010) allows censoring at any call attempt; however, this is not required for our purposes.

The full data likelihood is a function of the number of callbacks, the probability of contact, and the probability of an interview or noninterview given that an initial contact was made. For ease of exposition, we initially assume simple random sample (SRS) of size *n* households is selected from a population of size *N* households and later extend these results for complex survey sampling. Let π_g denote the proportion of the population in group *g* and let α_{ag} denote the probability a person in group *g* is contacted at call attempt *a* for *a* = 1,…,*K*. Let β_g denote the conditional probability that a person in group *g* is interviewed given the person is initially contacted at attempt *a*. Under this notation and assumptions, we can write the probability an individual has outcome (*a*,*d*,*g*) denoted by ρ_{adg} as

$$\begin{aligned}\rho_{adg} &= \pi_g \alpha_{ag} \beta_g \left[\prod_{t=1}^{a-1}(1-\alpha_{tg})\right], \text{ for } d=1 \\ &= \pi_g \alpha_{ag}(1-\beta_g)\left[\prod_{t=1}^{a-1}(1-\alpha_{tg})\right], \text{ for } d=2, \\ &= \pi_g \prod_{t=1}^{K}(1-\alpha_{tg}), \text{ for } d=3\end{aligned} \tag{1}$$

for *g* = 1,…,*L* and *a* = 1,…, *K*. The log-likelihood kernel the entire data set is

$$\ell = \sum_{a=1}^{K}\sum_{g=1}^{L} n_{a1g} \log \rho_{a1g} + \sum_{a=1}^{K}\sum_{d=2}^{3} n_{ad+} \log \sum_{g=1}^{L} \rho_{adg} \,. \tag{2}$$

Note that, in this likelihood, the response probabilities are summed over the *L* groups as $\sum_{g=1}^{L}\rho_{adg}$ for dispositions (2) and (3) because we assume that only the sums, $\sum_g n_{adg}$, are known for noninterviewed persons. Note that, by assumption, $n_{a3g} = 0$ for *a*=1,…,*K*-1 and thus (2) can be rewritten as

$$\ell = \sum_{a=1}^{K}\sum_{g=1}^{L} n_{a1g} \log \rho_{a1g} + \sum_{a=1}^{K} n_{a2+} \log \sum_{g=1}^{L} \rho_{a2g} + n_{K3+} \log \sum_{g=1}^{L} \rho_{a3g} \tag{3}$$

This likelihood contains many more parameters than can be estimated, and parameter restrictions must be imposed. An identifiable restricted model can be estimated by maximum likelihood estimation (MLE) via the expectation-maximization (EM) algorithm (Dempster, Laird, & Rubin, 1977) as described in Biemer, Chen, and Wang (2010). One such identifiable model assumes that $\alpha_{ag} = \alpha_g$ and $\beta_{ag} = \beta$ for all *a*=1,…,*K*, and *g*=1,…,*L*. Although this assumption is usually not satisfied for most survey data, the model can be a reasonable starting point for model selection. Under this assumption, the likelihood kernel in equation (2) can be rewritten as

$$\begin{aligned}\rho_{adg} &= \pi_g \alpha_g (1-\alpha_g)^{a-1}\beta, \text{ for } d=1 \\ &= \pi_g \alpha_g (1-\alpha_g)^{a-1}(1-\beta), \text{ for } d=2, \\ &= \pi_g (1-\alpha_g)^{K}, \text{ for } d=3\end{aligned} \tag{4}$$

for $g = 1,\ldots,L$ and $a = 1,\ldots, K$ with $2L+1$ parameters to be estimated: π_g, α_g for $g = 1,...,L$ and β.

The EM algorithm can be applied to the incomplete data to obtain MLEs of the model parameters by differentiating the full data likelihood with respect to each parameter, setting the derivatives to 0, and solving for the parameters in terms of the full table counts, n_{adg}. Expressions for the resulting estimators can be found in Biemer, Chen, and Wang (2010).

Additional α's can be introduced to account for possible variation in contact probabilities over call attempts. As an example, interviewers may obtain information about the sample persons at an earlier callback that can increase the probability of a future contact. Neighbors or other household members may suggest better times to call back or may indicate certain times of the day (like weekday afternoons) that are fruitless and should be avoided. Such information may change the contact probabilities at future attempts.

A model that specifies a separate contact probability for each attempt *a* and group *g* is identifiable. However, as Biemer, Chen, and Wang (2010) notes, adding too many α's can also cause problems of data sparseness. For this reason, it is usually more practical, efficient, and sufficient to specify unique contact probabilities for the first few attempts only. This is logical because after several visits to a household, interviewers are unlikely to acquire new information that would change the probability of contact on future attempts. Biemer, Chen, and Wang (2010) suggest three probabilities, α_{1g}, α_{2g}, and α_{3g}, for each group, where $\alpha_{ag} = \alpha_{3g}$, for a=3,…,K.

Biemer and colleagues formed 20 response propensity strata from the response propensities obtained from the current NSDUH response propensity model and then applied the callback model within each propensity stratum. Callback model estimates were obtained for each propensity stratum and then weighted together to obtain the callback model-adjusted estimate for the entire sample. Their results focused on the following four models;

- Model 0: $\pi_g, \alpha_{ag} = \alpha$, $\beta_{ag} = \beta$, for $g = 1,...,L, a = 1,...,K$
- Model 1: $\pi_g, \alpha_{ag} = \alpha_g$, $\beta_{ag} = \beta$, for $g = 1,...,L, a = 1,...,K$
- Model 2: same as Model 1, except, $\alpha_{1g}, \alpha_{ag} = \alpha_{2g}, a = 2,...,L$
- Model 3: same as Model 1, except, $\alpha_{1g}, \alpha_{2g}, \alpha_{ag} = \alpha_{3g}, a = 3,...,L$.

They also considered Model 1 after removing the constraint $\beta_{ag} = \beta$ denoted by Model 1′; i.e., this model specifies that β_{ag} may vary across the levels of *g*. However, this model produced poor results which they hypothesized was the result of errors in the callback data. This issue will be discussed in more detail subsequently.

Since age, race, sex, and ethnicity are obtain in the screener interview (some by proxy response) for all main interview persons, these characteristics are known and were used to construct gold standard estimates for the purposes of evaluating a given model's ability to adjust for nonignorable nonresponse bias.

3. Some Results from the NSDUH Analysis

The NSDUH interview process consists of a household screener used to enumerate household members and identify eligible respondents, followed by the selection of up to two members of the household for an interview. The NSDUH has unit nonresponse at two levels: the

screening interview and the main interview. Biemer, Chen, and Wang (2010) analyzed data for 85,034 respondents who were successfully screened and selected for an interview.

Table 1 shows the prevalence estimates for screener (*Y*) variables for the unadjusted model (Model 0) and the three nonresponse-adjusted models in our comparison. The estimate based upon full screener information is provided in the last column and will be considered the gold standard for this analysis. The estimates in the table are a weighted average for estimates over the 20 propensity strata and each model's ability to adjust for the missing variable using other variables in the model.

Table 1. Prevalence Estimates for Sex, Ethnicity, Age, and Race, by Model and Screener

Variable (*Y*)	Model 0	Model 1	Model 2	Model 3	Screener
Sex					
Males	47.1	48.0	47.5	47.4	48.2
Ethnicity					
Hispanics	14.3	16.0	14.7	14.5	13.8
Age					
12–17	13.1	12.9	13.0	13.1	10.4
18–25	15.7	17.0	15.8	15.9	13.2
26–34	14.4	14.2	14.5	14.5	14.2
35–49	25.1	24.8	25.3	25.3	26.2
50+	31.7	31.0	31.3	31.3	35.9
Race					
American					
Indian	1.1	1.1	1.2	1.1	0.9
Asian	3.1	3.7	3.2	3.2	4.7
Black	13.1	12.9	13.1	13.1	11.8
White	81.3	80.2	81.1	81.1	81.3
Multiple Race	1.4	2.1	1.5	1.5	1.3

The bias in an estimate can be computed by subtracting the gold standard (screener) estimate from the model estimate. These are shown in Table 2. Over the 20 strata, the bias in the Model 0 estimator is never very large for any of the four variables except for the oldest age category. Persons aged 50 and older are underestimated. The last two rows of Table 2 rank the models on their ability to adjust for nonignorable nonresponse by two criteria: absolute bias rank (denoted |Bias| Rank) and average absolute bias rank (denoted Avg |Bias| Rank). To compute |Bias| Rank, the absolute value of the biases in each row of the table were ranked from smallest

(rank of 1) to largest (rank of 6). The ranks were then averaged over the *Y* variables in the table for each model. By this criterion, the Model 3 does slightly better than the Models 1 and 2; however, none of the models is better than Model 0, which does use the callback information.

Note that the Model 0 estimates in Table 1 are adjusted for nonresponse since the propensity strata were formed using the NSDUH propensity model after removing the *Y* variable in the first column of Table 1. These results suggest that the callback information is not consistently effective in removing the "nonignorable" nonresponse bias induced by the absence of this *Y* variable in the NSDUH propensity model. One hypothesis as to why this is true is that errors in the LOE data cause biases in the callback model adjustments. To investigate this hypothesis further, an investigation of the quality of the LOE data was conducted.

Table 2. Nonignorable Bias for Sex, Ethnicity, Age, and Race, by Model

Variable (*Y*)	Model 0	Model 1	Model 2	Model 3
Sex				
Males	-1.05	-0.17	-0.69	-0.75
Ethnicity				
Hispanics	0.52	2.19	0.90	0.74
Age				
12–17	2.73	2.51	2.64	2.68
18–25	2.44	3.76	2.60	2.64
26–34	0.14	-0.02	0.29	0.23
35–49	-1.05	-1.35	-0.87	-0.93
50+	-4.26	-4.91	-4.65	-4.62
Race				
American Indian	0.21	0.24	0.27	0.23
Asian	-1.54	-0.95	-1.47	-1.49
Black	1.21	1.08	1.22	1.21
White	-0.02	-1.12	-0.23	-0.14
Multiple Race	0.15	0.76	0.23	0.20
|Bias| Rank	3.1	3.8	4.3	3.7
Avg |Bias| Rank	3.3	4.4	4.3	3.8

4. The Quality of the NSDUH Callback Data

These results suggest that the callback models are misspecified in some way because a well-specified model should be successful in removing nonignorable nonresponse bias. Based upon the experience from this application, we believe the problem does not stem from the lack of

fit of the models. For example, adding more α parameters to the model, while improving model fit, did not reduce the bias in the estimates of prevalence. In fact, an interesting phenomenon was observed regarding the β parameters. Models that allowed the β parameters to vary across the levels of Y markedly improved the model fit, yet the resulting prevalence estimates were substantially more biased. One possibility for these puzzling results is that data that are being used to model response propensity in the callback models (i.e., the LOE data) are flawed for the purpose of callback modeling.

The callback model assumes that the number call attempts are recorded accurately by the interviewers. The estimated response propensity for an individual is a function of the number of callbacks that were made on behalf of that individual to obtain an initial contact. Even small errors in the number of callbacks can cause bias in the estimates of response propensity. As a result, the nonresponse weighting adjustments that are functions of the estimated response propensities will be biased, resulting in either overcorrecting or undercorrecting the weights.

Because the quality of the callback data is an important question for this research, this question was addressed to some extent in this study by investigating the process used to collect the callback data. The purpose of this investigation was twofold. First, we wanted to learn about the accuracy of the callback data regarding the needs of the callback model. In particular, we sought information from field staff on how callback data are recorded and what factors might lead interviewers to record information on callbacks incorrectly. Second, we wished to obtain recommendations from the field staff for either improving the accuracy of entering callback data or making the process easier to carry out without reducing the accuracy or imposing additional burdens on the field staff.

First, an informal survey was conducted in September 2009 of all NSDUH interviewers that asked about their current reporting practices and how they would handle specific situations in the field. A total of 601 responses were obtained from 653 interviewers. In addition to the survey, two teleconference meetings were conducted with groups of field supervisors, regional supervisors, and regional directors during the same period.

A full report of the findings from this investigation appears in Wang and Biemer, 2010. However, the following bullet points summarize the main findings.

- Underreporting seems to be more frequent than overreporting of call attempts.
- Interviewers are prevented from overreporting because this practice can usually be discovered through timesheet reviews, and the consequences of intentionally falsifying these data are severe.
- Underreporting may occur because of pressures to keep a case "alive." If too many unproductive visits are recorded, the case may get closed out.
- Field staff wishing to avoid being perceived as not using time effectively may also underreport.
- Failure to report "drive-by" visits seems to be a primary cause of underreporting.
- Another frequent cause of underreporting is the failure to report attempts to interview multiple sample persons within the same household.
- The degree of underreporting can vary a lot by interviewer. This interviewer variance is very difficult to model in any nonresponse adjustment framework.
- Depending on the level of precision needed for callback modeling, potential changes range from modest (e.g., interviewer prompts when there are two interviews in a household) to extensive (e.g., new definitions of visits and procedures for recording visits).

This study concluded that the level of error in the callback data for callback modeling purposes is quite high and has the potential to seriously bias the results of the callback modeling approach to nonresponse adjustment. Our limited simulation study confirmed that even a small degree of underreporting (say 5%) is enough to appreciably bias the callback model estimates. But add to this the complications introduced by variation among supervisors, interviewers, and type of area in the level of underreporting and the situation becomes somewhat intractable to deal with through model enhancements alone. Rather, the preferred solution would seem to be improvements in the quality of the callback data for modeling purposes at its source.

Unfortunately, changing the field procedures currently in use for collecting these data and introducing additional quality checks to ensure accurate reporting of callbacks suitable for callback modeling would introduce additional burdens on interviewers. One concern is that increasing the burden associated with this task could draw the interviewer's attention away from crucial tasks such as contacting sample members, gaining their cooperation, and conducting the interview. A more prudent approach would be to wait until the NSDUH field systems are redesigned to take advantage of new technologies and other innovations. At that opportunity, a common purpose of the callback data for accuracy and other paradata should be considered a high priority.

5. Effects of Error in the LOE Data and its Effects on Parameter Estimation

Our investigations into the quality of the LOE data have led us hypothesize that errors in the number of call attempts may be, at least in part, responsible for the failure of the callback model to account for the nonignorable nonresponse in the analysis of the NSDUH screener variables. To further investigate this hypothesis, we simulated data sets for which the number of attempts to contact a sample member was underreported since that seemed to be the dominant error according to our field investigations. There are many ways to simulate error in the callback data. We used a very simple approach that we believe is still useful for examining estimation bias in the callback estimates resulting from underreported of call attempts.

Let u_{adg} denote the probability that the *a*th call attempt was not recorded for a unit in group *g* having disposition *d*. For purpose of demonstrating the effect of errors in the number of callbacks, we assume the probability an interviewer fails to record an attempt is the same for all attempts, which can be written as $u_{adg} = u_{a'dg} = u_{dg}$ for all a, a'. This assumption is probability an over simplification of the actual error mechanism because the probability failing to report a callback may well increase as the number of attempts increases. Nevertheless, we believe it is well-suited for purposes of illustrating the effect of errors on the estimates. In addition, additional complexity in the error structure should not change our primary conclusions. As a further simplification, we assume that underreporting occurs independently, both within cases and between cases.

These assumptions gives rise to the zero-truncated binomial distribution with parameters *a* and u_{dg} at each attempt, *a*. The distribution is zero-truncated because at least one attempt must be recorded for every case so the probability that the number of attempts is 0 for a case is 0. Let n'_{adg} denote the number of cases that are recorded to have *a* attempts with disposition *d* in group *g* where n_{adg} is the actual number. The expected value of n'_{adg} is, therefore, given by the following:

$$E(n'_{adg}) = \sum_{t=a}^{K} n_{tdg} \times b(t, a, u_{dg}) \tag{5}$$

where

$$b(t, a, u_{dg}) = \binom{t}{a} \frac{(1-u_{dg})^a u_{dg}^{t-a}}{1-u_{dg}^t} \tag{6}$$

To illustrate the effects of callback underreporting errors on the callback model parameters, we consider two levels of underreporting error—5 percent and 20 percent— and the biases resulting on the Model 1 parameter estimates. Many different of scenarios were considered that varied the error rates and models with essentially the same results. We further confine this demonstration to the case of a dichotomous outcome variable, *Y*, denoting users and nonusers, for example. The results reported here set $\pi_g = 0.15$ and $\alpha_1 = 0.4$, $\alpha_2 = 0.5$. Alternate values of these parameters were also considered, but did not change the conclusions. Six scenarios are reported as follows:

1. Low error for interviewed units only: $u_{1g} = 0.05$ and $u_{1g} = 0$ for g =1,2
2. Low error for refused only: $u_{2g} = 0.05$ and $u_{2g} = 0$ for g =1,2
3. Low error for all sample persons: $u_{dg} = 0.05$ for d =1,2,3 and g = 1,2
4. High error for interviewed units only: $u_{1g} = 0.20$ and $u_{1g} = 0$ for g =1,2
5. High error for refused only: $u_{2g} = 0.20$ and $u_{2g} = 0$ for g =1,2
6. High error for all sample persons: $u_{dg} = 0.20$ for d =1,2,3 and g = 1,2

Each of these six scenarios was simulated for two alternative levels of interview probabilities, as shown in Table 3: equal probabilities (i.e., $\beta_1 = \beta_2 = 0.8$, shown in the top half of the table) and unequal probabilities (i.e., $\beta_1 = 0.75$, $\beta_2 = 0.85$, shown in the bottom half of the table). For the former scenarios, Model 1 was fit to the simulated data; i.e., the model specifying equal contact and interview probabilities for all callbacks, group heterogeneous contact probabilities, and group homogeneous interview probabilities. Model 1′ was fit to the latter scenarios; i.e., Model 1 except interview probabilities was allowed to vary across groups.

The top half of Table 3 provides compelling evidence that the effects of underreporting errors on the estimates of are π_1 fairly small for Model 1, even at the 20 percent error level. This is good news because unbiased estimation of π_1 is key. The biases are more substantial for α_1 and α_2 and β_1 and β_2, but these biases appear to have little biasing effect on π_1.

The bottom half of the table tells a different story. When β_1 and β_2 differ and Model 1′ is fit, underreporting errors can have a substantial effect even for small levels of error. For example, when the underreporting rate is only 5% for interviewed persons and 0% for refusals, the bias is 0.179 – 0.15 = 0.064, a relative bias of about 43%. Likewise, when the underreporting rate is 5% for refusals and noncontacts and 0% for interviews, the bias changes direction and is

smaller, but is still substantial with a relative bias of about -16%. Interestingly, setting the error rate to 5% for the entire sample has little effect on the bias. These effects are magnified as the error rates increase. At a 20% error rate, the relative bias is 84% and -25%, respectively, for the two types of errors. The bias is negligible, even at 20%, when the same error rate applies to all sample members.

Table 3. Estimates of Simulation Population Parameters under Model 1 for Twelve Scenarios ($\pi = 0.15,\ \alpha_1 = 0.40,\ \alpha_2 = 0.50$ **)**

	$\beta_1 = \beta_2 = 0.80$					
Scenario	u_{dg}	$\hat{\pi}_1$	$\hat{\alpha}_1$	$\hat{\alpha}_2$	$\hat{\beta}_1$	$\hat{\beta}_2$
Interviewed	0.050	0.151	0.413	0.516	0.800	0.800
Ref/NC	0.050	0.150	0.404	0.504	0.800	0.800
All	0.050	0.150	0.417	0.519	0.800	0.800
Interviewed	0.200	0.152	0.453	0.565	0.800	0.800
Ref/NC	0.200	0.148	0.413	0.514	0.800	0.800
All	0.200	0.150	0.474	0.582	0.800	0.800
	$\beta_1 = 0.75,\ \beta_2 = 0.85$					
Interviewed	0.050	0.179	0.418	0.519	0.630	0.880
Ref/NC	0.050	0.126	0.340	0.500	0.895	0.826
All	0.050	0.151	0.417	0.519	0.747	0.851
Interviewed	0.200	0.276	0.475	0.582	0.407	0.998
Ref/NC	0.200	0.113	0.400	0.510	0.990	0.814
All	0.200	0.152	0.4741	0.582	0.741	0.852

Finally, similar results were obtained for overreporting error using a similar model for misreporting error. Due to space, those tables are not included here.

6. Discussion

Biemer, Chen, and Wang (2010) applied a number of callback models, including the models described in Section 2, to data from the NSDUH in an attempt to adjust for nonignorable nonresponse bias in the reported NSDUH estimates. They reported that small biases remained in the estimates of π after implementing the callback models. They also reported that callback models that allowed the β parameters to differ by the values of *y* (for e.g., Model 1′ above) produced very poor results. Our simulation offers some explanation for their findings.

The simulations conducted for this study suggest that underreporting seems to have a much lesser effect on Model 1 than on Model 1′. This is consistent with the empirical results from Biemer, et al (2010) that found the estimates from Models 1, 2, and 3 to be slightly biased. However, when the β-parameters were allowed to vary across the values of y (as for Model 1′), the bias was considerable. The simulations further suggest that the most damaging types of errors are errors that depend upon a unit's final disposition. When the errors are unrelated to a unit's final disposition, the bias was negligibly small.

Errors that disposition dependent may not be uncommon is face to face surveys. For example, cases that are ultimately interviewed may have received more intense followup by interviewers. When the number of followup attempts is larger, interviewers may tend to underreport them more frequently. On the other hand, most of the cases in the refused/noncontacted group are refusals since the NSDUH as a very low noncontact rate after 15 call attempts. Once a case finally refuses, further callbacks are suspended and, thus, the chance of underreporting callbacks is reduced. Therefore, the results of this simulation study provide a possible explanation for the findings in Biemer, et al (2010).

Still, these findings are by no means definitive and are limited by several factors. First, our callback error model is admittedly overly-simplified. The actual error generation mechanism is likely to be considerably more complex. For example, it is likely that there is interviewer variance in the underreporting error. Our field study of callback error found some evidence of this. Moreover, underreporting, besides being related to the final disposition, may also be related to other factors such as neighborhood characteristics, time of year, and type of unit (apartment, guarded entrance, etc.). Such errors could heighten the effects noted in the simulations. Finally, this study only considered the simplest callback models (Models 1 and 1′). The effects of errors on models that allow both contact and interview probabilities to vary by callback have not been considered. Possibly these models are more affected by callback data errors.

Further study of the impact of callback errors on the model estimates is needed to deal with these complexities. However, ultimately solutions are needed to fully realize the potential of callback data in the nonresponse modeling process. One obviously solution is to collection better callback data; however, this would like entail greater emphasis on accurately recording call attempts which could divert interviewer attention away from more critical activities. Another solution may be to incorporate a callback error model into the callback response model. The callback error model discuss in this paper may be a first step in that direction.

7. References

Biemer, P.P., Chen, P., & Wang, K. (2010). *Using Level of Effort Paradata in Nonresponse Adjustments with Application to Field Surveys*. Internal RTI Report.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). "Maximum likelihood from incomplete data via the EM algorithm." *Journal of the Royal Statistical Society, B*, 1–38.

Wang, K. H., & Biemer, P. (2010). "The accuracy of interview paradata: Results from a field investigation." *Proceedings of the American Association for Public Opinion Research*, Chicago, IL (May).