Journal of Official Statistics, Vol. 17, No. 2, 2001, pp. 295–320

Nonresponse Bias and Measurement Bias in a Comparison of Face to Face and Telephone Interviewing

Paul P. Biemer[1]

This article provides a study design and analytic methodology for evaluating and comparing the quality of survey data in the case of face to face and telephone interviewing. Under the proposed design, the mode differences are decomposed into measurement bias and nonresponse bias components. The measurement bias is estimated from the interview-reinterview data using latent class analysis which simultaneously estimates the true prevalence rates and the classification error rates for the measures of the characteristics of interest. Nonresponse bias for the face to face survey is estimated from a followup survey of the face to face nonrespondents. Nonresponse bias for the telephone survey is estimated using an error-corrected estimator of the true prevalence rate. The methodology is illustrated using data from a special study conducted for the U.S. National Health Interview Survey (NHIS). Although the study population is limited to Texas and California, the analysis provides new insights regarding the nature of the mode effects for these two interview modes while illustrating an innovative design for assessing mode bias.

Key words: Latent class models; nonsampling error; National Health Interview Survey; health questions.

1. Introduction

There have been numerous studies in the survey methods literature that compare the estimates obtained from a random-digit-dialing (RDD) telephone survey with those of a face to face survey of the same population. (See, for example, De Leeuw and Van der Zouwen (1998) and Groves (1989) for comprehensive reviews of the literature through the mid-eighties.) An important objective in these studies is to determine whether survey estimates produced from the modes are equal and, if not, to determine which mode is better in the sense of giving smaller total biases for the items of interest. These so-called mode comparison studies usually involve a split-sample experiment where one random subsample is assigned to the face to face mode and the other to the telephone mode. The data are collected using essentially the same questionnaire and interviewing procedures. The estimates are compared and any significant differences between estimates of the same population parameter are attributed to biases arising from one or both interview modes.

For some mode comparison studies, there may be very little information available on the magnitudes of the mode biases. However, if information is available on the direction

[1]Research Triangle Institute, Box 12194, Research Triangle Park, NC 27709, U.S.A. E-mail: ppb@rti.org.
Acknowledgements: The author would like to acknowledge the substantial contributions of Don Malec to this research as well as the assistance of Van Parsons in the review of this report. Helpful comments were also received from Monroe Sirken and James Massey.

of the biases for both modes, the less-biased mode can sometimes be identified. For example, for questions on sensitive topics such as alcohol consumption and drug use, it is sometimes assumed that the mode of interview providing the higher prevalence of the sensitive behavior is less biased since the tendency in the population would be to under-report sensitive behaviors. Sykes and Collins (1988) report on several split-sample experiments in Great Britain that used this approach to compare estimates collected by centralized telephone and face to face interviews for a number of sensitive topics. They found slightly higher reporting of these behaviors in the telephone mode and concluded that the telephone mode was less biased than the face to face mode.

An obvious shortcoming of this approach is the reliance on the often dubious assumption that higher reporting of sensitive behavior reflects more accurate reporting. For example, in a study of sexual behavior, Turner et al. (1998) found that unmarried males tend to exaggerate their sexual activity while unmarried females tends to underreport it. The assumption of a consistent tendency among all population groups to underreport extramarital sexual activity could lead to a wrong conclusion as to which mode is better in this instance. Another danger of this approach is that it ignores a number of other mode-related biases that confound the direct comparison of the estimates and may behave quite differently than the measurement bias. As we discuss subsequently, the mode effect includes not only measurement bias but also bias due to nonresponse and other factors. Thus, attributing the difference between the estimates from two modes of interview to measurement bias is often inappropriate. Biemer (1988) discusses other limitations of the so-called more-is-better approach for determining the preferred mode.

An alternative method for assessing mode effects is to incorporate features in the study design that yield direct estimates of the bias associated with each mode. These features usually involve comparing the survey data to external data that are assumed to be more accurate than the survey estimates and, thus, can be used as a gold standard for comparing the two modes. For example, in an income study in Denmark, Körmendi (1988) estimated the telephone and face to face mode biases using income data from the tax authorities. Also, De Leeuw and Van der Zouwen (1988) conducted a meta-analysis of 31 telephone or face to face mode comparison studies. About a third of these studies estimated the biases associated with each mode using comparisons of the survey data to some type of administrative data which could be considered as a gold standard.

Biemer (1988) cites several limitations of the use of external data such as administrative records for evaluating survey accuracy. Some of these are: the unavailability of external data for all the variables of interest in an evaluation; differences in the definitions of the variable in the survey and those available on records; differences in the time periods covered by the survey and the external data; and ambiguities in matching survey respondents to administrative records. These limitations may lead to biases in the estimated mode effects that could result in misleading conclusions regarding which is the more accurate mode.

In addition, since the total bias attributable to a mode of interview is actually the sum of biases from a number of error sources, interpreting the results of total bias comparisons is risky and potentially misleading. For example, two key components of the total bias are measurement bias and nonresponse bias. As we shall see in this article, these component biases may be off-setting; that is, the sign of nonresponse bias may be opposite to the sign

of the measurement bias resulting in a small overall bias from the two sources combined. Thus, a small estimated total bias could conceal substantial component biases. If the survey were repeated and achieved a higher response rate, employed better interviewers, or incorporated other improvements that altered the mix of biases, the total bias could actually be higher. However, without separately estimating the biases associated with each error source, the analyst may be unaware of this situation.

In this article, we implement an approach that does not rely on assumptions regarding the direction of the bias or the accuracy of the validation data. Our approach is a model-based approach to the estimation of the major component biases comprising the mode effects. The focus is on two modes of interview – face to face (F-to-F) interviewing and telephone interviewing, where the latter is performed in a centralized facility using Computer Assisted Telephone Interviewing (CATI). Using a somewhat novel study design and modeling approach, we develop the statistical methodology for separately estimating the nonresponse and measurement biases associated with each mode and demonstrate the methodology on data from a mode comparison study conducted for the U.S. National Health Interview Survey (NHIS).

In Section 2, we describe the study design and data collection methodology for estimating the mode bias components. Section 3 describes the statistical model and estimation approach for estimating the biases. In Section 4, we present the results of the study and, finally, in Section 5, the key findings and their implications are summarized.

2. The Study Design

The study design will be described in the context of the 1994 NHIS since this survey was the focus of our evaluation. However, the design is quite general and can be applied to most surveys that can be conducted either by F-to-F or CATI. I begin with a short description of the NHIS.

The NHIS, conducted by the National Center for Health Statistics (NCHS), is a continuing survey of the civilian noninstitutionalized population 17 years of age or older living in the U.S. Its purpose is to provide national data on the incidence of illness and injury, the prevalence of diseases and impairments, the extent of disability, the use of health services, and other health-related topics. The annual sample size is approximately 49,000 eligible occupied households containing around 132,000 individuals. A full description of the NHIS sample design for 1982–1996 may be found in Massey, Moore, Parsons, and Tadros (1989).

From 1982 to 1996, the survey questionnaire consisted of two parts: 1) a set of health and demographic items (known as the Core questionnaire) and 2) a supplement, referred to as the Healthy People 2000 (HP2000) Supplement, which was designed to provide information to assess national progress toward the President's Healthy People 2000 Program goals. Our study focused on the CATI and F-to-F mode effects for a subset of items on the HP2000 Supplement. The study was conducted in two states – California and Texas – that were selected because their NHIS sample sizes were large enough to support the analyses planned for the study. In each state, the five data collection components listed in Table 1 were conducted.

Data for the F-to-F survey component were collected by the U.S. Census Bureau using

Table 1. Study design components and mode of interview

Study component	Description	Mode
NHIS HP2000 Supplement	HP2000 Supplement information collected as part of the usual national NHIS sample in each state	F-to-F
NHIS Nonresponse Followup	Survey of all NHIS HP2000S nonrespondents that could be contacted by telephone	CATI
NHIS Reinterview	Reinterview survey administered to approximately a half-sample of the NHIS respondents in each state	CATI
CATI HP2000 Supplement	HP2000 Supplement information collected for an independently selected RDD sample in Texas and California	CATI
CATI Reinterview	Reinterview administered to approximately a 25 percent sample of RDD HP2000 respondents in each state	CATI

paper and pencil interviewing methods as part of the regular NHIS sample, while the other components were collected in the Research Triangle Institute's centralized CATI facility. The HP2000 Supplement was conducted with a randomly selected person 17 years of age or older living in the NHIS sample household.

The CATI survey was conducted coincidentally with the 1994 NHIS survey during the months of February 1994 through November 1994. The universe for the CATI survey is the civilian noninstitutional population 17 years of age or older in California and Texas who resided in housing units having a working telephone within the unit. The sample for the CATI survey used an RDD sampling scheme described in Biemer and Akin (1994) that resulted in a simple random sample of all telephone households within the two states. Within each eligible household, a person was randomly selected from all eligible household respondents using the Trodahl-Carter within household selection scheme (see Trodahl and Carter 1964).

To maintain comparability and consistency in all data collection components, questions in the F-to-F survey, the CATI surveys, the reinterview surveys, and the nonresponse followup surveys were identical except for any wording changes necessary to maintain the same reference periods for all components and to account for the inability to use visual aids such as flash cards in the telephone mode.

Each week, the U.S. Census Bureau transmitted the NHIS HP2000 Supplement interviewing results to RTI who used this information to select the NHIS reinterview and NHIS nonresponse followup samples. Except for the transmittal operation, there were no other changes to the U.S. Census Bureau's regular NHIS procedures. For the reinterview survey sample, the CATI interview was replicated for a random sample of HP2000 Supplement respondents. A self-response respondent rule was imposed on all reinterviews; i.e., only the original respondent could respond to the reinterview. All reinterviews were conducted within 10 to 28 days of the original interview. The reinterview questions and procedures followed a test-retest design; i.e., the reinterview attempted to

replicate the essential survey conditions of the original interview to the greatest extent possible.

The F-to-F survey nonresponse followup operation targeted all nonrespondents to the HP2000 Supplement in Texas and California who could be reached by telephone, including whole unit nonrespondents as well as HP2000 Supplement only nonrespondents. Nonresponse followup interviews were conducted within two weeks of the final F-to-F survey interview attempt using the same respondent rule and procedures as the main survey components.

A nonresponse followup of CATI nonrespondents was also attempted but ultimately discontinued due to an unacceptably low contact and interview rate. Fortunately, this component is not required in our analysis of CATI nonresponse bias, as we shall see subsequently.

Table 2 provides the number of interviews and response rates for all components of the study. The CATI response rate was computed by estimating the number of unresolved units that are in-scope following Hidiroglou, Drew, and Gray (1993), i.e.,

$$\mathrm{RR} = \frac{C}{I + aU} \tag{1}$$

where C is the number of completed interviews, I is the number of all in-scope units, U is the number of unresolved numbers – i.e., ring-no-answer (RNA), answering machine numbers, and other non-contact cases – and a is an estimate of the proportion of unresolved numbers that are in-scope. This estimate, which was computed as the proportion of *resolved* numbers that are in-scope, was 42.7 percent overall, 44.1 percent for California, and 41.2 percent for Texas. The response rates for California and Texas were 56.8 percent and 64.3 percent, respectively, with an overall response rate of 60.3 percent.

For the CATI reinterview survey, the average conditional response rate was 73.7 percent, somewhat less in California (72.2 percent) than in Texas (75.2 percent). For the F-to-F survey, the response rate was higher in California than in Texas: 82.4 percent compared with 79.5 percent, with an overall rate of 81.3 percent. The F-to-F reinterview response rate was lower in both sites than for the CATI survey: 66.4 percent and 74.3 percent for California and Texas, respectively. Note, however, that the combined interview-reinterview response rates were still substantially higher for the F-to-F survey than for the CATI survey. For the CATI survey, the combined rates were 41 and 48 percent for California and Texas, respectively, while for the F-to-F survey the corresponding rates were 55 and 59 percent, respectively.

Table 2. Sample yields and response rates (percent) by study component

	California		Texas		Total	
	n	RR	*n*	RR	*n*	RR
F-to-F Survey	1,614	82.4	910	79.5	2,524	81.3
F-to-F Nonresponse followup	105	29.9	109	47.6	214	37.1
F-to-F Reinterview	1,072	66.4	675	74.3	1,747	69.3
CATI Survey	2,112	56.8	2,122	64.3	4,234	60.3
CATI Reinterview	653	72.2	712	75.2	1,365	73.7

3. Estimation and Modeling of Mode Effects

3.1 Notation and definitions

To eliminate the confounding of the CATI and F-to-F comparisons by nontelephone household coverage bias, the subsequent analyses will be confined to telephone households only in each mode. Further, our analysis initially considers the bias in a state-level estimator and then combines the state-level estimates for a study-area summary estimator. Therefore, unless otherwise noted, the notation will pertain to estimates for a particular state.

Let π denote the proportion of the telephone population in a state who possess some characteristic, y. For the CATI survey, let η_C denote the proportion of the target population in the state in the nonresponse subpopulation (i.e., η_C is the expected nonresponse rate for the CATI survey). Then, it can be shown (see, for example, Cochran 1977) that

$$\pi = \eta_C \pi_{N,C} + (1 - \eta_C)\pi_{R,C} \tag{2}$$

where $\pi_{N,C}$ is the prevalence of the characteristic among nonrespondents and $\pi_{R,C}$ is the prevalence of the characteristic among respondents.

Let $p_{R,C}$ and $p_{R,F}$ denote the usual, design-based estimators of π based upon data from the CATI survey and F-to-F survey, respectively. Here the subscript R emphasizes that the estimators pertain to the respondent populations for each mode. Then, the nonsampling error bias in $p_{R,C}$ is

$$\begin{aligned} \text{Bias}\,(p_{R,C}) &= \mathrm{E}(p_{R,C} - \pi) \\ &= \eta_C(\pi_{R,C} - \pi_{N,C}) + \mathrm{E}(p_{R,C} - \pi_{R,C}) \\ &= \eta_C \Delta_C + M_C, \text{ say.} \end{aligned} \tag{3}$$

In the last line, the first term after the equality is the nonresponse bias component and the second term is the measurement bias, M_C. Note that the nonresponse bias is the product of two terms – the expected CATI nonresponse rate and the difference between respondent and nonrespondent population proportions which shall be denoted by Δ_C.

Analogously, it can be shown that

$$\begin{aligned} \text{Bias}\,(p_{R,F}) &= \mathrm{E}(p_{R,F} - \pi) \\ &= \eta_C(\pi_{R,F} - \pi_{N,F}) + \mathrm{E}(p_{R,F} - \pi_{R,F}) \\ &= \eta_F \Delta_F + M_F, \text{ say.} \end{aligned} \tag{4}$$

where $\pi_{R,F}$ is the prevalence of the characteristic among F-to-F respondents and $\pi_{N,F}$ is the prevalence among F-to-F nonrespondents, η_F is the expected F-to-F nonresponse rate, Δ_F is the expected difference between F-to-F respondents and nonrespondents, and M_F is the F-to-F measurement bias.

3.2 Estimation of measurement biases

For a dichotomous characteristic, y, let π denote the prevalence of y in the population, φ the false positive probability, θ the false negative probability, and let p denote the estimate of π from the survey. From a well-known result (see, e.g., Biemer and Stokes 1991, p. 501),

$$\mathrm{E}(p) = \pi(1 - \theta) + (1 - \pi)\varphi \tag{5}$$

and the bias in $\hat{\pi}$ is $M = \mathrm{E}(p) - \pi$ or

$$M = -\pi\theta + (1 - \pi)\varphi \tag{6}$$

Let $\hat{\theta}$ and $\hat{\varphi}$ denote consistent estimates of θ and φ, respectively. It will be shown subsequently (see Equation 15) that a consistent estimator of M is

$$\hat{M} = p - \hat{\pi} \tag{7}$$

where

$$\hat{\pi} = \frac{p - \hat{\varphi}}{(1 - \hat{\theta} - \hat{\varphi})} \tag{8}$$

In this section, we estimate the measurement biases, M_{C} and M_{F}, defined in the last section by employing an estimation approach referred to as latent class analysis (LCA). Through LCA, we will obtain estimates of φ and θ from the test-retest reinterview data and then use these estimates to correct the survey estimate, p, for measurement error as in (8). The resulting estimator, $\hat{\pi}$, is a consistent estimator for π and, therefore, consistent estimators of M_{C} and M_{F} can be formed similarly by subtracting from $p_{\mathrm{R,C}}$ and $p_{\mathrm{R,F}}$, their corresponding measurement error corrected estimators.

Clogg (1995) provides a comprehensive review of LCA and discusses its many applications in social science research and psychometrics. LCA is a model-based estimation approach which uses the patterns of inconsistency in the interview-reinterview table to obtain estimates of the classification error probabilities associated with measurement processes. The model assumptions made for classical LCA are similar to those made for other analysis of test-retest reinterview data; for example, it assumes independent errors between the interview and reinterview responses (see, for example, Borhnstedt 1983). Our approach is similar to that used by Hui and Walter (1980) and Sinclair and Gastwirth (1996).

To simplify the discussion, we first consider the case of a dichotomous response variable and later extend the methodology to polytomous categorical variables. Let μ_k denote the true classification for the kth unit in either the CATI or F-to-F sample and let y_{tk} denote the observed classification for the unit at time t where $t = 1$ denotes the interview and $t = 2$ denotes the reinterview. For dichotomous measures, $\mu_k = 1$ if the kth unit is a true ''yes'' and $\mu_k = 0$ if a true ''no.'' Let g denote a grouping variable such as a pre or a post-stratification variable, experimental treatment, or other explanatory variable for the measurement error analysis where $g = 1, \ldots, L$. Define the following parameters:

$\pi_g =$ $\Pr(\mu_k = 1 | g)$; i.e., the true prevalence rate for target population members in group g,

$\theta_{gt} =$ $\Pr(y_{tk} = 0 | \mu_k = 1, g, t)$; i.e., the false negative probability for group g at time t,

$\varphi_{gt} =$ $\Pr(y_{tk} = 1 | \mu_k = 0, g, t)$; i.e., the false positive probability for group g at time t, and

$P_{ij|g} =$ $\Pr(y_{1k} = i, y_{2k} = j | g)$ is the probability of a unit in group g being classified in the (i, j) cell of the interview-reinterview table.

The false negative probability is the probability a true positive (i.e., person possessing the characteristic) is erroneously classified as a negative (i.e., not possessing the characteristic).

The false positive probability is the probability a true negative is erroneously classified as a positive.

An assumption that is inherent in LCA models is the so-called local independence assumption where it is assumed that $\Pr(y_{1k} = y, y_{2k} = y' | \mu_k) = \Pr(y_{1k} = y | \mu_k) \Pr(y_{2k} = y' | \mu_k)$ for $y \neq y'$. This is equivalent to the assumption of between trial independence of response errors in test-retest reliability studies. In our application, this assumption may be violated if respondents try to recall their original responses and simply repeat them in the reinterview rather than independently arriving at their responses. Unfortunately, it is not possible to test this assumption with only interview-reinterview data, but Hagenaars (1988) provides a test when three or more measures are available.

For our study, the field procedures were designed to maintain local independence. For example, the timing of the reinterview is critical. Too soon after the initial interview and the local independence assumption is at risk, too late and another key assumption may be violated: that of equal classification errors between occasions (to be described below). We believe the 10 to 28 day window following the interview for conducting the reinterview provides adequate protection for this assumption.

Another assumption that is inherent in test-retest data analysis is the assumption of equal error distributions for the two measurements (also referred to as ''parallel'' measurements in Borhnstedt 1983). When the second interview is intended to replicate the first using identical procedures, skill-levels of the interviewer, mode of interview, etc., it is reasonable to assume that the classification error probabilities, θ_g and φ_g, apply to both the interview and reinterview measurement processes. Indeed, replicating the CATI interview was a key feature of the reinterview survey design. The assumption may be violated if the essential survey conditions for the reinterview survey are considerably different than those for the original survey; for example, if the reinterview is conducted by less experienced interviewers or if a long period of time elapses between the two interviews. In the latter situation, recall error may lead to an increase in classification error for some items in the reinterview violating the assumption of parallel measures.

In our application, the assumption of equal measurement error distributions for interview and reinterview is plausible for the CATI component since both interviews were conducted by CATI under identical survey conditions. However, it may be violated for the F-to-F component since those reinterviews were also conducted by CATI. As we will see subsequently, this feature does not limit our ability to estimate the classification errors separately for each mode. Clogg (1995) discusses methods for testing this assumption when only two measurements are available.

To illustrate the simplest case the LCA likelihood, we initially assume local independence and parallel measures; however, the latter assumption will be relaxed in our analysis. It therefore follows that

$$
\begin{aligned}
P_{i=1,\, j=1 |\, g} &= \pi_g(1-\theta_g)(1-\theta_g) + (1-\pi_g)\varphi_g\varphi_g \\
P_{i=1,\, j=0 |\, g} &= \pi_g(1-\theta_g)\theta_g + (1-\pi_g)\varphi_g(1-\varphi_g) \\
P_{i=0,\, j=1 |\, g} &= \pi_g\theta_g(1-\theta_g) + (1-\pi_g)(1-\varphi_g)\varphi_g \\
P_{i=0,\, j=0 |\, g} &= \pi_g\theta_g\theta_g + (1-\pi_g)(1-\varphi_g)(1-\varphi_g)
\end{aligned} \tag{9}
$$

which can be written more concisely as

$$P_{ij|g} = \pi_g(1-\theta_g)^{i+j}\theta_g^{2-i-j} + (1-\pi_g)\varphi_g^{i+j}(1-\varphi_g)^{2-i-j} \tag{10}$$

where i (or j) $= 1$ for a positive response and 0 for a negative response.

Thus, the cell probabilities, $P_{ij|g}$, which are obtained directly from the observed data, can be expressed in terms of the unknown true prevalence rate and the latent classification parameters. Likewise, the likelihood for the interview-reinterview table can be expressed in terms of these parameters and, under certain conditions, the parameters can be estimated using maximum likelihood estimation techniques. Let γ_g denote the proportion of the target population in group g, then the likelihood of the $2 \times 2 \times G$ interview-reinterview table for G groups is

$$\boldsymbol{L}(\boldsymbol{P}_{ijg}) = C \prod_g \prod_i \prod_j \gamma_g P_{ij|g}^{n_{ijg}} \tag{11}$$

where C is a constant and n_{ijg} is the number of units in cell (i,j) of the interview-reinterview table for group g, and the $P_{ij|g}$ are the probability cell probabilities given by (10).

To apply this model for estimating the classification error probabilities for CATI and F-to-F interviewing, we define the grouping variable, g, to be a combination of two variables: the state variable s where $s =$ TX or CA and the mode variable m where $m =$ C, for CATI, or F, for F-to-F. Thus, our design consists of four groups (i.e., $G = 4$) as follows: $(s,m) =$ (TX, C), (TX, F), (CA, C), and (CA, F). Before the model for $P_{ij|g}$ can be applied for these groups some alternative assumptions to those made for (11) are needed.

As previously noted, the CATI main survey as well as the reinterview survey for the CATI sample were conducted using the same telephone facility, the same staff of interviewers and supervisors, and the same questions. Additionally, the reinterviews for the F-to-F survey were conducted by CATI simultaneously with the reinterviews for the CATI survey using the same staff, procedures, etc. It is therefore reasonable to assume that the error parameters for the CATI survey and the two reinterview surveys are equal as follows:

1. $\theta_{\text{TX,C,1}} = \theta_{\text{TX,C,2}} = \theta_{\text{TX,F,2}} = \theta_{\text{TX,tel}}$, say,
   $\varphi_{\text{TX,C,1}} = \varphi_{\text{TX,C,2}} = \varphi_{\text{TX,F,2}} = \varphi_{\text{TX,tel}}$, say,
   $\theta_{\text{CA,C,1}} = \theta_{\text{CA,C,2}} = \theta_{\text{CA,F,2}} = \theta_{\text{CA,tel}}$, say, and
   $\varphi_{\text{CA,C,1}} = \varphi_{\text{CA,C,2}} = \varphi_{\text{CA,F,2}} = \varphi_{\text{CA,tel}}$, say.

In addition, in the CATI facility, no distinction was made between Texas cases and California cases as far as the interviewing process was concerned. Again, since the same CATI interviewers, supervisors, procedures, and questions were used in both states, it is plausible to assume that the CATI classification errors are equal for the two states as well; i.e., we further assume

2. $\theta_{\text{TX,tel}} = \theta_{\text{CA,tel}} = \theta_{\text{tel}}$, say, and
   $\varphi_{\text{TX,tel}} = \varphi_{\text{CA,tel}} = \varphi_{\text{tel}}$, say.

Assumptions (1) and (2) in combination with the local independence assumption constitute the assumptions for the base measurement error model for the study. Note that we initially do not assume that the error parameters for the F-to-F survey are equal to

the CATI survey parameters; rather their equality is a key research question to be tested in our analysis. If the test of equality is rejected for a survey item, we will conclude that the mode effect due to measurement bias is significant for the item. Further, our initial LCA model does not assume that the F-to-F parameters are equal across the two states; although this assumption also can be tested. Thus, the F-to-F survey parameters – $\theta_{\mathrm{TX,F,1}}$, $\varphi_{\mathrm{TX,F,1}}$, $\theta_{\mathrm{CA,F,1}}$, and $\varphi_{\mathrm{CA,F,1}}$ – are unrestricted in the initial LCA model within the interval [0, 1]. Hereafter we will drop the subscript ''1'' on these parameters to simplify the notation.

Having dealt with the restrictions on the classification error parameters, we turn our attention to the prevalence parameters, π_g. Although we have confined our analysis to telephone households in both states, the respondent population for each mode differs. This is evidenced by the differential response rates. For example, the combined CATI and reinterview survey response rate is 45.0 per cent while for the F-to-F survey it is 56.3 per cent (see Table 2). Therefore, to allow for these differences in the responding populations by mode, we specify separate parameters for the prevalence rates of the characteristic under the two modes of interview for Texas and California and denote these as $\pi_{\mathrm{CA,C}}$, $\pi_{\mathrm{CA,F}}$, $\pi_{\mathrm{TX,C}}$, and $\pi_{\mathrm{TX,F}}$.

Thus, the base model has 10 parameters to be estimated while the number of cells in the $2\times2\times4$ group by interview by reinterview table is 16. Subtracting four degrees of freedom for γ_g ($g=1,\ldots,4$) leaves 2 degrees of freedom for testing model fit. As shown in Goodman (1974), nonnegative model degrees of freedom is not a sufficient condition for estimability in LCA so the identifiability of the model parameters must be confirmed in the estimation process. The joint likelihood for the $2\times2\times4$ is still given by (11) except now the form of $P_{ij|g}$ is changed to reflect assumptions (1) and (2). Similar to (10), it can be shown that for the CATI survey in state s, the probability of an observation in (i,j) of the interview-reinterview table is

$$P_{ij|\mathrm{CATI},s}=\pi_{s,C}(1-\theta_{\mathrm{tel}})^{i+j}\theta_{\mathrm{tel}}^{2-i-j}+(1-\pi_{s,C})\varphi_{\mathrm{tel}}^{i+j}(1-\varphi_{\mathrm{tel}})^{2-i-j} \tag{12}$$

and for the F-to-F survey in state s it is

$$P_{ij|\mathrm{F},s}=\pi_{s,\mathrm{F}}(1-\theta_{s,\mathrm{F}})^{i}(1-\theta_{\mathrm{tel}})^{j}\theta_{s,\mathrm{F}}^{1-i}\theta_{\mathrm{tel}}^{1-j}+(1-\pi_{s,\mathrm{F}})\varphi_{s,\mathrm{F}}^{i}\varphi_{\mathrm{tel}}^{j}(1-\varphi_{s,\mathrm{F}})^{1-i}(1-\varphi_{\mathrm{tel}})^{1-j} \tag{13}$$

In what follows, we will use maximum likelihood estimation to estimate the ten parameters for binary response variables: $\pi_{s,\mathrm{C}}$, $\pi_{s,\mathrm{F}}$, θ_{tel}, φ_{tel}, $\theta_{s,\mathrm{F}}$, and $\varphi_{s,\mathrm{F}}$ for $s=$ TX and CA.

Extending the latent class analysis for response variables with three or more categories is straightforward and will not be given here. However, the model assumptions are equivalent to assumptions (1) and (2) in that classification probabilities for the CATI and reinterview survey are assumed to be equal and these rates are the same in Texas and California.

Now consider a variable with K (≥2) response categories and let $\mathbf{p}'=(p_1,p_2,\ldots,p_K)$ denote the K observed proportions corresponding to the K categories for either the CATI or F-to-F survey. For example, when $K=2$, p_1 is either $p_{\mathrm{R,C}}$ or $p_{\mathrm{R,F}}$ and p_2 is $1-p_1$. Let a_{ij} be the probability that an observation which truly belongs to the ith category is assigned to the jth category and let π_i denote the true proportion in the population in the ith category. Then

$$\mathbf{p}=\mathbf{A}'\boldsymbol{\pi}_R \tag{14}$$

where $\boldsymbol{\pi}_R = (\pi_1, \ldots, \pi_K)'$ and $\mathbf{A} = [a_{ij}]$ is the $K \times K$ matrix with elements a_{ij} (Rao and Thomas, 1991).

The classification probability matrix $\mathbf{A}$ can be estimated using the latent class modeling approach just described and applied to $\mathbf{p}$ to obtain an estimator of $\boldsymbol{\pi}$ that is corrected for measurement error. Letting $\hat{\mathbf{A}}'$ denote the transpose of this estimator, it follows that an estimator of $\boldsymbol{\pi}$ corrected for measurement error is

$$\hat{\boldsymbol{\pi}}_R = (\hat{\mathbf{A}}')^{-1}\mathbf{p} \tag{15}$$

Thus, $\mathbf{p} - \hat{\boldsymbol{\pi}}_R$ is a consistent estimator of the measurement bias in $\mathbf{p}$.

To illustrate, suppose $K = 2$ and let $a_{11} = (1 - \hat{\theta})$, $a_{12} = \hat{\theta}$, $a_{21} = \hat{\varphi}$, $a_{22} = (1 - \hat{\varphi})$, $p_1 = p$ and $p_2 = (1 - p)$. It can easily be verified that substitution of these expressions in (15) yields (8).

3.3. Estimation of nonresponse bias

We first consider the estimation of the F-to-F nonresponse bias component, Δ_{F}, in the dichotomous case. Let $p_{\mathrm{N,F}}$ denote the estimator of the proportion, π, for the F-to-F nonrespondents based upon the nonresponse followup survey. The usual estimator of the difference between respondent and nonrespondent populations is the simple difference between respondent and nonrespondent prevalence estimates given by

$$d_{\mathrm{R,N}}(\text{F-to-F}) = \hat{\pi}_{\mathrm{R,F}} - \hat{\pi}_{\mathrm{N,F}} \tag{16}$$

(see, for example, Groves and Couper 1998, p. 79). Note, however, that both terms on the right in (16) are subject to measurement error. The prevalence estimator for the respondent population is subject to error arising from the face to face mode and the estimator for the nonrespondent population is subject to error arising from the CATI mode since the nonresponse followup was conducted by CATI. Moreover, the latter estimator is itself also subject to nonresponse bias since the nonresponse rate for the nonresponse followup was approximately 63 percent (cf. Table 2). In what follows, we will correct both terms in (16) for measurement error using a correction factor as in (15) that employs LCA estimates of measurement error from the reinterview survey. However, we have no valid means for correcting (16) for the bias due to the followup survey nonresponse. Fortunately, we can show that the nonresponse bias in (16) and the nonresponse bias in the corresponding estimator for the CATI component, Δ_{C}, are equal. These biases will cancel when the two estimates are contrasted and thus comparisons of CATI and F-to-F nonresponse biases will be unbiased although the separate estimates may be biased.

To correct (16) for measurement error, we apply (15) to each term on the right in (16) where now $\mathbf{A}$ in (15) is the appropriate classification matrix estimated from the LCA and $\mathbf{p}$ is the vector of estimates for the F-to-F respondents or nonrespondents, as appropriate. For a questionnaire item with K categories, let $\mathbf{p}_{\mathrm{F,R}}$ denote the K-vector of observed proportions for the F-to-F survey and let $\mathbf{p}_{\mathrm{F,N}}$ denote the corresponding K-vector for the telephone followup of F-to-F nonrespondents. Let $\mathbf{A}_{\mathrm{F}}$ and $\mathbf{A}_{\mathrm{tel}}$ denote $K \times K$ misclassification probability matrices for the face to face and telephone interview modes, respectively. Since the persons in the nonresponse followup sample were interviewed by CATI and persons in the F-to-F sample were interviewed by face to face methods, we will use $\mathbf{A}_{\mathrm{tel}}$ to correct for measurement error in $\mathbf{p}_{\mathrm{F,N}}$ and $\mathbf{A}_{\mathrm{F}}$ to correct for measurement

error in $\mathbf{p}_{F,R}$. Let $\boldsymbol{\Delta}_{\mathbf{F}}$ denote the K-vector of the components $\boldsymbol{\Delta}_{\mathbf{F}}$ in (4) corresponding to the K categories of the item. Then, applying (15), an estimator of $\Delta_{\mathbf{F}}$ corrected for measurement error is

$$\hat{\boldsymbol{\Delta}}_{\mathbf{F}} = (\hat{\mathbf{A}}'_{\mathbf{F}})^{-1}\mathbf{p}_{\mathbf{F,R}} - (\hat{\mathbf{A}}'_{\mathbf{tel}})^{-1}\mathbf{p}_{\mathbf{F,N}} \tag{17}$$

where $\hat{\mathbf{A}}_{\mathbf{F}}$ and $\hat{\mathbf{A}}_{\mathbf{tel}}$ are estimators of $\mathbf{A}_{\mathrm{F}}$ and $\mathbf{A}_{\mathrm{tel}}$, respectively, from the latent class analysis in the previous section.

Note that using $\hat{\mathbf{A}}_{\mathbf{tel}}$ in (17) to correct for the measurement error in the CATI nonresponse followup assumes that the nonrespondents who respond in the followup survey have the same measurement error distributions as the CATI survey respondents. This assumption may not hold if, for example, the nonresponse followup respondents ''satisfice'' (i.e., put less effort into providing good responses) to a greater extent than the CATI survey respondents. Also, since the followup interviews are conducted at the same time as the reinterviews, the time difference between interviews may change the measurement error distribution, particularly for questions that require recall. Notwithstanding these limitations, the assumption seems plausible since the CATI nonresponse followup interviews were carried out under conditions which were otherwise identical to the CATI main survey conditions. Unfortunately, these data provide no means to test this assumption since conducting reinterviews with the nonresponse followup respondents is not practical.

Now consider the estimation of Δ_{C} in (3), i.e., the nonresponse component for the CATI survey. Since a nonresponse followup study was not conducted for the CATI sample, a direct estimator of Δ_{C} cannot be constructed from the data as was done for Δ_{F}. However, it is possible to estimate this component indirectly as follows.

Since both the measurement bias and nonresponse bias can be estimated for the F-to-F survey, an estimator of the true prevalence rate for the whole telephone population can be constructed by summing the respondent and nonrespondent measurement error corrected estimators; thus an estimator of $\boldsymbol{\pi}$ constructed from the measurement error corrected, area sample data is

$$\hat{\boldsymbol{\pi}} = (1 - \eta_{\mathrm{F}})(\hat{\mathbf{A}}'_{\mathbf{F}})^{-1}\mathbf{p}_{\mathbf{F,R}} + \eta_{\mathrm{F}}(\hat{\mathbf{A}}'_{\mathbf{tel}})^{-1}\mathbf{p}_{\mathbf{F,N}} \tag{18}$$

where as in (4) η_{F} is the NHIS HP 2000 supplement nonresponse rate. To obtain an estimator of the CATI nonresponse bias, we subtract (18) from our measurement error corrected estimator of the CATI respondent prevalence estimator. Let $\mathbf{p}_{\mathrm{C,R}}$ denote the estimator from the CATI survey of the true CATI prevalence parameters, $\boldsymbol{\pi}_{\mathrm{R,C}}$, a K-vector corresponding to the K-categories of the item. In analogy to (17), let Δ_{C} denote the K-vector of $\boldsymbol{\Delta}_{\mathrm{C}}$-components as in (3) for the K categories of the item. Then a consistent estimator of $\boldsymbol{\Delta}_{\mathrm{C}}$ is

$$\hat{\boldsymbol{\Delta}}_{\mathbf{C}} = \eta_{\mathrm{C}}^{-1}[(\hat{\mathbf{A}}'_{\mathbf{tel}})^{-\mathbf{1}}\mathbf{p}_{\mathbf{C,R}} - \hat{\pi}] \tag{19}$$

Alternative estimators of the F-to-F and CATI biases are possible. However, the proposed estimators have the property that the difference between the two total bias estimates is equal to the difference, $\mathbf{D}$, between the two design-based estimators. That is,

$$\mathbf{D} = \mathbf{p}_{\mathbf{C}} - \mathbf{p}_{\mathbf{F}} = \hat{\mathbf{B}}_{\mathbf{C}} - \hat{\mathbf{B}}_{\mathbf{F}} \tag{20}$$

Table 3. Summary of bias components and estimators

Line	Bias Vector	Description	Estimator
1	$\mathbf{M_C}$	Measurement bias for CATI survey estimator, $\mathbf{p}_{C,R}$	$\hat{\mathbf{M}}_\mathbf{C} = \mathbf{p}_{\mathbf{C,R}} - \hat{\boldsymbol{\pi}}_{\mathbf{C,R}}$ where $\hat{\boldsymbol{\pi}}_\mathbf{C} = (\hat{\mathbf{A}}_\mathbf{tel})^{-1}\mathbf{p}_{\mathbf{C,R}}$
2	$\mathbf{M_F}$	Measurement bias for F-to-F survey estimator, $\mathbf{p}_{F,R}$	$\hat{\mathbf{M}}_\mathbf{F} = \mathbf{p}_{\mathbf{F,R}} - \hat{\boldsymbol{\pi}}_{\mathbf{F,R}}$ where $\hat{\boldsymbol{\pi}}_\mathbf{C} = (\hat{\mathbf{A}}_\mathbf{F})^{-1}\mathbf{p}_{\mathbf{F,R}}$
3	$\eta_C\boldsymbol{\Delta}_\mathbf{C}$	Nonresponse bias for CATI survey estimator, $\mathbf{p}_{C,R}$	$\eta_C\hat{\boldsymbol{\Delta}}_\mathbf{C} = [(\hat{\mathbf{A}}'_\mathbf{tel})^{-1}\mathbf{p}_{\mathbf{C,R}} - \hat{\boldsymbol{\pi}}]$ where $\hat{\boldsymbol{\pi}}$ is given by (18)
4	$\eta_F\boldsymbol{\Delta}_\mathbf{F}$	Nonresponse bias for F-to-F survey estimator, $\mathbf{p}_{F,R}$	$\eta_\mathrm{F}\hat{\boldsymbol{\Delta}}_\mathbf{F} = [(\hat{\mathbf{A}}'_\mathbf{F})^{-1}\mathbf{p}_{\mathbf{F,R}} - \hat{\boldsymbol{\pi}}]$ where $\hat{\boldsymbol{\pi}}$ is given by (18)
5	$\mathbf{B_C}$	Total bias for CATI survey estimator, $\mathbf{p}_{C,R}$	$\hat{\mathbf{B}}_\mathbf{C} = \hat{\mathbf{M}}_\mathbf{C} + \eta_C\hat{\boldsymbol{\Delta}}_\mathbf{C} = \mathbf{p}_{\mathbf{C,R}} - \hat{\boldsymbol{\pi}}$
6	$\mathbf{B_F}$	Total bias for F-to-F survey estimator, $\mathbf{p}_{F,R}$	$\hat{\mathbf{B}}_\mathbf{F} = \hat{\mathbf{M}}_\mathbf{F} + \eta_\mathrm{F}\hat{\boldsymbol{\Delta}}_\mathbf{F} = \mathbf{p}_{\mathbf{F,R}} - \hat{\boldsymbol{\pi}}$

where $\hat{\mathbf{B}}_\mathbf{C}$ and $\hat{\mathbf{B}}_\mathbf{F}$ denotes the sum of the measurement bias and nonresponse bias estimates for CATI and F-to-F, respectively. Thus, the usual estimator of the mode effect, $\mathbf{D}$, is an unbiased estimator of the difference in total biases between the two modes, by our definitions. In addition, it is easily shown that the biases in the estimates given by (17) and (19) are equal so that $\eta_C\hat{\boldsymbol{\Delta}}_\mathbf{C} - \eta_\mathrm{F}\hat{\boldsymbol{\Delta}}_\mathbf{F}$ is a consistent estimator of $\eta_C\boldsymbol{\Delta}_\mathbf{C} - \eta_\mathrm{F}\boldsymbol{\Delta}_\mathbf{F}$, thus negating the effect of nonresponse to the nonresponse followup survey on comparisons of nonresponse bias by mode. To see this, note that (17) is equivalent to

$$\hat{\boldsymbol{\Delta}}_\mathbf{F} = \eta_\mathrm{F}^{-1}[(\hat{\mathbf{A}}'_\mathbf{F})^{-1}\mathbf{p}_{\mathbf{F,R}} - \hat{\boldsymbol{\pi}}] \tag{21}$$

and, thus, from (19) we see that the difference between the two nonresponse biases does not depend upon $\hat{\boldsymbol{\pi}}$.

Table 3 provides a summary of the measurement bias, nonresponse bias, and total bias estimators for design-based estimators of telephone population estimators for CATI and F-to-F surveys. We will refer to this table in the next section to identify the estimators used in the analysis.

In our analysis, we report on the standard errors of the estimated total biases and the difference between the total biases. The standard error of the total bias was computed by combining the estimates of standard errors of the measurement error and nonresponse bias components, ignoring any covariance between the components, which should be trivial. Estimates of the standard errors were derived from the maximum likelihood standard errors provided by the PANMARK software and do not take into account sample design effects. However, for both the RDD sample and the area sample, design effects for the items in our analysis were in the range 0.6 to 1.5 and approximately 1.0 on average. Further details on the computation of standard errors for the estimates can be found in Biemer (1997).

4. Results

4.1 Fitting the measurement error models

Data on more than 200 survey items were collected in both the main surveys for the CATI, the F-to-F, the two reinterview surveys, and the nonresponse followup survey. However,

the results reported here are confined to a subset of 14 variables selected for their substantive interest to the U.S. National Center for Health Statistics. This set of questions include items such as the number of smoke detectors in the home, smoking behavior, blood pressure, doctor visits, and firearms in the home. Table 4 is a listing of all these variables with a short description of each.

In Section 3, we showed how LCA models can be fit to the interview-reinterview data in order to estimate the measurement error biases for the items in Table 4 using the formulas summarized in Table 3. The basic model in our analysis, given in (12) and (13), specifies separate measurement error parameters for CATI and F-to-F interviewing, thus allowing hypotheses regarding the presence of a mode effect due to measurement error to be tested. This model also assumes state-specific classification error probabilities for the F-to-F mode while for the CATI mode the error parameters are assumed to be equal across states (cf. assumptions 1 and 2 above). In what follows, the basic model will be referred to as the Mode and State Effects Model and two alternative models that add restrictions to the model will also be considered. Again, we initially describe the models for the simple case of a dichotomous response variable and then extend the set-up for polytomous variables.

The first alternate model is appropriate for testing whether measurement error biases for the F-to-F survey differ by state. If they do not differ, a more parimonious model for mode effects can be used in the bias estimation process, yielding potentially more precise estimates. This model, which will be referred to as the No State Effect Model, adds the following assumption to assumptions (1) and (2) above:

3. $\theta_{\mathrm{CA,F}} = \theta_{\mathrm{TX,F}} = \theta_{\mathrm{F}}$, say and $\varphi_{\mathrm{CA,F}} = \varphi_{\mathrm{TX,F}} = \varphi_{\mathrm{F}}$, say.

In other words, the model assumes no difference between California and Texas for the F-to-F survey classification errors parameters.

The third and last model we will consider is also the most parsimonious model. This model specifies a fourth assumption that restricts the error parameters for CATI and F-to-F modes to being equal. Referred to as the No Mode or State Effects Model, the model adds the assumption:

4. $\theta_{\mathrm{F}} = \theta_{\mathrm{tel}} = \theta$, say and $\varphi_{\mathrm{F}} = \varphi_{\mathrm{tel}} = \varphi$, say.

That is, the model assumes there are no differences in the error parameters either by mode of interview or by state.

We fitted each of these three models separately for the variables in Table 2, using the PANMARK software (Van de Pol, Langeheine, and De Jong 1991). The data were weighted for the F-to-F and CATI selection probabilities and rescaled so that the sum of the cell counts for each variable equaled the number of completed interview-reinterview pairs for the variable. Thus, the results of statistical tests reflect the actual number of observations in the sample weighted to telephone population distributions. The PANMARK software, like all available software for latent class analysis, assumes simple random sampling. Therefore, the standard errors of the LCA estimates do not account for the complex survey design of the NHIS and, consequently, the NHIS estimated standard errors and model p-values may be understated.

To determine which of the three models best fits the data for a particular survey item, we used the model selection rule suggested by Lin and Dayton (1997), which consists of two criteria. First, the p-value associated with the likelihood ratio chi-squared, L^2, for the model must be at least 0.05, which indicates that the model is reasonably consistent with the data. Second, for models satisfying the first criterion, the best model is the one having the smallest Akaike Information Criterion (AIC) value. The AIC value (Akaike, 1987) is given by $-2\log(L) + 2npar$, where $npar$ denotes the number of parameters in the model. Since AIC is a trade-off between the fit of the model (the $-2\log(L)$ term) and the number of parameters in the model, it may be interpreted as a measure of model parsimony. Thus, the Lin-Dayton criteria aim to identify the most parsimonious model that is consistent with the data. In this article, variables that did not satisfy the first criterion for at least one of the three models considered were excluded from the latent class analysis.

Table 5 provides the results of fitting the three LCA models. The best model by our selection criteria is highlighted in the table. Note that only three variables failed the first criterion: A8a (Anyone Smoke Inside Home?), D8a (Time Since Blood Pressure Checked?), and E2b (Time Since Last Checkup?). These variables were eliminated in the subsequent analysis. For the other variables, the p-values ranged from 0.05 to 0.69. Among the 11 remaining variables, two exhibit significant mode and state effects, five exhibit no significant state effects, and four exhibit neither significant mode effects nor state effects. Thus, significant mode effects were observed for 7 of the 11 variables. As a final check on the models selected by the above criteria, a likelihood ratio L^2 test was performed for each model to directly test for the presence of mode and state effects. In every case, the results of these tests were consistent with the AIC-based selection method.

4.2 Estimates of the bias components

Next, we use the results of the latent class analysis to estimate measurement bias and nonresponse bias for both modes of data collection and both states using the general formulas in Table 3. Then we combine the mode and nonresponse bias components to estimate the total bias for CATI and F-to-F survey estimators as shown in lines 5 and 6 in Table 3. Since the standard errors of the state-level estimates are quite large, we only report the estimates for the combined two-state area which were computed as a simple average of the two state estimates.

Table 6 provides a summary of these results and is divided into two parts: the results of the CATI survey on the left side of the table and the results of the F-to-F on the right. For each survey, the first numerical column is the weighted survey estimate unadjusted for measurement error and nonresponse bias, the second is the estimator of the measurement bias using Table 3, lines 1 and 2, and the third is the estimator of the nonresponse bias using Table 3, lines 3 and 4, and the nonresponse rates (i.e., the $\hat{\eta}$'s) from Table 2. Finally, the fourth and fifth numerical columns under the survey type are the total bias estimates using Table 3, lines 5 and 6 and the model-based estimate of its standard error.

Table 7 compares the estimates from the CATI survey, labeled $p_{R,C}$, and the F-to-F, labeled $p_{R,F}$. From (20), the difference between the two estimates, D, in the table is equal to the difference of the two mode biases. Thus, the standard error in the table is the standard error of this difference corrected for the NHIS complex sample design.

Table 4. Questions selected for the latent class analysis

No.	Question Label	Question Wording	Response Categories
A2A	NO. OF SMOKE DETECTORS?	How many smoke detectors are installed in your home?	None, One, Two, Three or more
A8A	ANYONE SMOKE?	Does anyone who lives in your home smoke cigarettes, cigars, or pipes anywhere inside your home?	Yes, No
B1	LIFETIME SMOKING?	Have you smoked at least 100 cigarettes in your entire life?	Yes, No
B2	SMOKE LAST YEAR?	Around this time last year, were you smoking cigarettes every day, some days, or not at all?	Every Day, Some Days, Not at All
B3A	SMOKING NOW?	Do you now smoke cigarettes every day, some days, or not at all?	Every Day, Some Days, Not at All
B3B	TIME SINCE QUIT?	How long has it been since you quit smoking cigarettes?	1 Yr. or Less, 1–3 Years, 3+ Years
B5	LAST YEAR STOPPED SMOKING?	During the past 12 months, have you stopped smoking for one day or longer?	Yes, No
B7	WANT TO QUIT?	Would you like to completely quit smoking cigarettes?	Yes, No
D1	HIGH BP?	Have you ever been told by a doctor or other health professional that you had hypertension, sometimes called high blood pressure?	Borderline/Pregnancy, Yes, No

D8A	TIME SINCE BP CHECKED?	About how long has it been since you had your blood pressure checked by a doctor or other health professional?	Never, 6 Mos. or Less, 1 Year or Less, More than 1 Year
E1	GENERAL HEALTH?	Would you say your health in general is excellent, very good, good, fair, or poor?	Excellent/VG, Good/Fair, Poor
E2A	REASON FOR LAST DOCTOR VISIT?	What was the reason for your last visit to a medical doctor or other health professional? Was it for a new problem, followup of a previous problem, a general physical exam, (an ob/gyn checkup, related to pregnancy,) or something else?	New Problem, Previous Problem, General Exam, Other
E2B	TIME SINCE CHECKUP?	About how long has it been since your last general physical exam or routine checkup by a medical doctor or other health professional? Do not include a visit about a specific problem.	Less than 1 Year, 1–2 Years, 2–3 Years, 3+ Years
G1	FIREARMS IN THE HOME?	Are any firearms now kept in or around your home? Include those kept in a garage, outdoor storage area, truck, or car	Yes, No

Table 5. Results of fitting the three LCA models (shading indicates model selected.)

	Model for State and Mode Effects				Model for No State Effect				Model for No Mode or State Effects			
Question	X^2	d.f.	*p*	AIC	X^2	d.f.	*p*	AIC	X^2	d.f.	*p*	AIC
A2a	15	12	0.24	18,806	30.80	24	0.16	18,797	55.70	36	0.02	18,798
A8a*	10.6	2	0.00	12,401	15.10	4	0.00	12,402	60.80	6	0.00	12,443
B1	3.1	2	0.21	13,498	13.10	4	0.01	13,504	44.40	6	0.00	13,531
B2	5.1	6	0.53	7,038	7.60	12	0.82	7,028	18.60	18	0.42	7,027
B3a	13.3	6	0.04	6,818	19.20	12	0.08	6,812	36.40	18	0.01	6,817
B3b	9.3	6	0.16	3,015	14.10	12	0.29	3,008	16.80	18	0.54	2,998
B5	1.9	2	0.39	1,954	6.20	4	0.18	1,954	7.50	6	0.28	1,952
B7	0.69	2	0.71	1,606	5.90	4	0.21	1,607	12.20	6	0.06	1,609
D1	3.9	6	0.69	13,501	13.30	12	0.35	13,498	21.60	18	0.25	13,494
D8a*	28.7	12	0.00	16,121	48.80	24	0.00	16,118	88.90	36	0.00	16,134
E1	10.3	6	0.11	16,085	20.40	12	0.06	16,083	48.60	18	0.00	16,099
E2a	18.6	12	0.10	22,262	26.50	24	0.33	22,246	83.50	36	0.00	22,279
E2b*	30.6	12	0.00	20,052	51.60	24	0.00	20,049	77.80	36	0.00	20,052
G1	3.7	2	0.16	12,608	6.60	4	0.16	12,607	38.70	6	0.00	12,635

* Model selection criteria not satisfied. Items were deleted from the analysis.

Table 6. Comparison of RDD and NHIS bias components: both states combined

		RDD (value × 100%)					NHIS (value × 100%)				
Item no.	Item cat.	$p_{R,C}$	$\hat{M}_{RDD}$	$\eta_{RDD}\hat{\Delta}_{RDD}$	$\hat{B}_C$	*s.e.* ($\hat{B}_C$)	$p_{R,F}$	$\hat{M}_{NHIS}$	$\eta_{NHIS}\hat{\Delta}_{NHIS}$	$\hat{B}_F$	*s.e.* ($\hat{B}_F$)
A2A	None	11.64	0.16	−1.88	−1.72	0.58	14.63	0.70	0.57	1.26	0.72
	One	36.47	−3.15	−2.44	−5.59	0.91	41.31	−2.04	1.29	−0.75	0.88
	Two	30.79	0.46	1.48	1.94	0.82	27.21	−0.54	−1.09	−1.63	0.75
	Three+	21.11	2.53	2.84	5.37	0.69	16.86	1.88	−0.77	1.12	0.66
B1	Yes	43.76	−1.36	−3.36	−4.72	0.85	44.72	−2.98	−0.79	−3.77	0.79
	No	56.24	1.36	3.36	4.72	0.85	55.29	2.98	0.79	3.77	0.79
B2	Every Day	33.94	3.95	−2.51	1.44	1.34	36.84	2.57	1.77	4.34	1.82
	Some Days	14.28	−6.62	3.83	−2.78	1.44	10.77	−4.85	−1.44	−6.29	1.52
	Not At All	51.79	2.67	−1.33	1.34	1.51	52.40	2.28	−0.33	1.95	1.76
B3A	Every Day	33.60	1.64	3.04	4.68	1.27	36.15	5.69	1.53	7.23	1.22
	Some Days	11.87	0.14	−1.61	−1.47	1.14	10.05	−2.74	−0.56	−3.29	0.90
	Not at All	54.54	−1.78	−1.42	−3.20	1.34	53.81	−2.96	−0.98	−3.93	1.21
B3B	1 Year or <	12.22	−0.80	2.54	1.75	1.06	9.68	−0.53	−0.27	−0.80	0.97
	1–3 Year	10.68	−0.86	−0.63	−1.49	−1.38	9.42	−0.89	−1.87	−2.76	1.04
	3+ Years	77.10	1.65	−1.92	−0.26	1.58	80.92	1.42	2.13	3.56	1.29
B5	Yes	51.84	−11.03	0.32	−10.71	2.44	49.19	−10.47	−2.90	−13.37	2.12
	No	48.16	11.03	−0.32	10.71	2.44	50.82	10.47	2.90	13.37	2.12
B7	Yes	73.07	3.17	4.36	7.53	2.19	72.34	1.64	5.16	6.81	1.83
	No	26.94	−3.17	−4.36	−7.53	2.19	27.66	−1.64	−5.16	−6.81	1.83
D1	Brdln/Preg	2.19	−2.82	0.24	−2.57	0.29	1.99	−2.56	−0.20	−2.77	0.34
	Yes	21.72	−1.14	2.15	1.01	0.63	21.26	−1.22	1.77	0.55	0.69
	No	76.10	3.95	−2.39	1.57	0.65	76.75	3.78	−1.57	2.22	0.70
E1	Excel/VG	52.72	10.08	0.17	10.26	0.98	60.39	13.62	4.30	17.93	0.93
	Good/Fair	45.58	−9.47	2.71	−6.76	1.00	37.19	−10.81	−4.34	−15.15	0.95
	Poor	1.72	−0.61	−2.87	−3.49	0.27	2.42	−2.81	0.03	−2.78	0.34
E2A	New Prob	25.43	−5.13	−2.95	−8.09	2.14	28.02	−6.29	0.79	−5.50	2.96
	Prev. Prob	22.51	3.98	−3.87	0.12	1.55	29.26	5.34	1.53	6.87	2.14
	Gen. Exam	31.26	0.77	4.08	4.86	3.45	28.25	1.97	−0.13	1.84	4.13
	Other	20.81	0.38	2.73	3.11	3.15	14.48	−1.02	−2.19	−3.22	3.78
G1	Yes	35.70	−0.48	−3.17	−3.65	0.71	37.15	−2.10	−0.09	−2.19	0.86
	No	64.31	0.48	3.17	3.65	0.71	62.85	2.10	0.09	2.19	0.86

Table 7. The differences between RDD and NHIS estimates (telephone households only) and their standard errors

Item no.	Item cat.	$\hat{\pi}$ (RDD)	$\hat{\pi}$ (NHIS)	Difference (D)	$s.e.$ (D)
A2A	None	11.64	14.63	−2.99**	1.11
	One	36.47	41.31	−4.85**	1.31
	Two	30.79	27.21	3.58**	1.17
	Three+	21.11	16.86	4.26**	1.09
B1	Yes	43.76	44.72	−0.96	1.24
	No	56.24	55.29	0.96	1.24
B2	Every Day	33.94	36.84	−2.90	1.82
	Some Days	14.28	10.77	3.51**	1.21
	Not At All	51.79	52.40	−0.61	1.93
B3A	Every Day	33.60	36.15	−2.55	1.81
	Some Days	11.87	10.05	1.82	1.15
	Not At All	54.54	53.81	0.73	1.89
B3B	1 Year or <	12.22	9.68	2.55	1.56
	1–3 Year	10.68	9.42	1.27	1.46
	3+ Years	77.10	80.92	−3.82	1.98
B5	Yes	51.84	49.19	2.66	3.27
	No	48.16	50.82	−2.66	3.27
B7	Yes	73.07	72.34	0.72	2.45
	No	26.94	27.66	−0.73	2.45
D1	Brdln/Preg	2.19	1.99	0.20	0.33
	Yes	21.72	21.26	0.46	1.08
	No	76.10	76.75	−0.65	1.10
E1	Excel/VG	52.72	60.39	−7.67**	1.29
	Good/Fair	45.58	37.19	8.39**	1.29
	Poor	1.72	2.42	−0.71*	0.34
E2A	New Prob	25.43	28.02	−2.59*	1.11
	Prev. Prob	22.51	29.26	−6.76**	1.12
	Gen. Exam	31.26	28.25	3.02*	1.31
	Other	20.81	14.48	6.33**	0.97
G1	Yes	35.70	37.15	−1.46	1.31
	No	64.31	62.85	1.46	1.31

* significant at $\alpha = 0.05$, ** significant at $\alpha = 0.01$, *** significant at $\alpha = 0.001$

The results in Table 7 demonstrate that neither mode of interview is better for all the variables in the analysis. F-to-F interviewing has lower total bias for A2a while CATI is preferred for the general health question, E1. Both modes have large biases for E2a, the question on the last doctor visit. For most of the other variables, the differences between the total biases are not significant. In terms of the magnitude of the total bias estimates, the F-to-F total bias exceeds the CATI survey total bias 19 times out of 31 item response categories across the 11 variables.

The mode differences are more apparent when the individual bias components are considered. The absolute value of the CATI nonresponse bias exceeds (although not always significantly) that of the F-to-F nonresponse bias for 23 of the 31 items. However, the absolute value of the F-to-F measurement bias exceeds that of the CATI bias for 17 of the items.

As a summary comparison, we compared the absolute value of the bias components and

Table 8. Summary of bias estimates by mode

	Average measurement bias	Average nonresponse bias	Average total bias
CAT1	3.12	2.39	4.13
F-to-F	3.64	1.59	4.84

total bias averaged across all 31 response categories in the analysis and Table 8 summarizes these results. As expected the average nonresponse bias is substantially higher for the CATI survey than the F-to-F survey: 2.39 versus 1.59. Note that the measurement bias tends to be much larger, on average, than nonresponse bias in both modes. Moreover, this bias is slightly higher in the F-to-F mode than in the telephone mode. The total bias averaged across all items in the survey is very comparable for the two modes, however. This is partly due to the measurement bias for CATI being smaller in absolute value than for the F-to-F mode, and partly because the measurement and nonresponse biases for CATI are offsetting (14 items for CATI and 8 items for F-to-F). This table provides evidence of the somewhat surprising result that measurement bias can exceed nonresponse bias for many items in a survey even when response rates are relatively low.

Now consider the total bias patterns for specific questions. For question A2a, number of smoke detectors, significant mode differences are observed for all four categories (see Table 7). From Table 6, the sources of the differences are evident. For example, there appear to be too few reports of one smoke detector households and too many reports of three or more smoke detector households for the telephone mode due to both telephone measurement error and telephone nonresponse. The F-to-F mode exhibits very small biases for this item.

With regard to the questions on cigarette smoking, there are significant biases for most of the questions (Table 7). Both surveys tend to underestimate the proportion of persons who have smoked at least 100 cigarettes in their lifetimes (Question B1), primarily due to measurement bias. For Question B2, we found no significant classification error differences by mode in Table 5. However, the total bias is significant and is due primarily to observing too few individuals in the ''Some days'' category. This appears to be due to the combination of nonresponse bias and measurement bias for F-to-F for this category.

For Question B3a on current smoking behavior, both surveys exhibit bias toward more frequent smoking. This bias is slightly higher for the F-to-F, but not significantly so. The biases for B3b (Time Since Quit Smoking) are small for both surveys and not significantly different. However, the F-to-F total bias is significant for two categories; ''1–3 years'' and ''3+ years.'' Both surveys underestimate the number of persons who quit smoking during the past year (Question B5) and overestimate the number of persons who want to quit smoking (Question B7). For the former question, the bias is due primarily to measurement error, which is large and not significantly different between modes (as noted in Table 5). For the latter question, nonresponse and measurement bias contribute about equally to the CATI bias, while for F-to-F the bias is primarily due to measurement error. Note from Table 7, however, that except for one category in item B2, the proportions are not significantly different for these items.

Now considering the remaining health questions, also both surveys exhibit total biases

which are significant and considerable. For D1, the question regarding being told about high blood pressure by a physician, both surveys tend to overestimate the ''no'' category. Also, surveys also show a tendency to overestimate the proportion holding opinions of excellent health (Question E1), primarily as a result of measurement bias. However, the bias toward more optimistic responses is significantly higher for the NHIS. Note from Table 6 that the difference between the total biases for the general health question is significant. Both surveys display a considerable amount of classification error for question E2A – reason for the last visit to the doctor. (Also significant differential total bias from Table 7.)

For the question on firearms (G1), which is a politically sensitive question in the U.S., both surveys exhibit a slight bias toward underestimating the proportion of the population that keep firearms in the household, primarily as a result of nonresponse bias for the CATI survey and measurement bias for the F-to-F mode.

Additional analysis of these data is provided in the full report on the study (see, Biemer 1997). In that report, the indexes of inconsistency (or test-retest reliability coefficients) for F-to-F and CATI were compared for the characteristics in Table 4. As shown in that report, since the F-to-F interviews and reinterviews were conducted by different modes, the usual estimates of test-retest reliability are biased, and must be corrected for the measurement error differences between telephone and F-to-F interviewing. For the F-to-F survey reinterview, the correction can be easily made using the estimated response probability matrix, $\hat{\mathbf{A}}_{\mathbf{tel}}$. Both the corrected and uncorrected estimates of reliability revealed that, for a majority of the characteristics in Table 4, the F-to-F reliabilities are significantly smaller than the corresponding reliabilities for CATI, indicating larger measurement error in the F-to-F mode. Thus, the test-retest reliability and the bias analysis provide consistent statements regarding the measurement error in the two modes of interview.

In their article, Cannell et al. (1987) speculated on the reasons the telephone interview protocol performs better than the F-to-F mode for many NHIS characteristics. They conjecture that the CATI's lower measurement bias may not be related to the mode itself but may reflect better performance by telephone interviewers. For our study, the CATI interviewers worked in a centralized location, permitting ready interaction with each other and the project staff. This was exploited extensively in the study to improve interviewer performance and build interviewer enthusiasm for the survey objectives. Thus, as Cannell et al. state, ''the better reporting performance perhaps reflects highly motivated, well-trained, and supervised interviewers.'' They further conclude that ''telephone interviews can match personal interviews on the quality of the data and have the potential for producing better data, at least for some survey topics.'' We agree with their assessment of the reasons for improved data quality using the telephone and have provided additonal evidence in support of their conclusion.

5. Summary and Conclusions

In this article, we considered the mode biases in estimates of parameters derived from an RDD CATI survey and from a comparable area sample face to face survey. The difference between the estimates can be attributed to four error sources: measurement error, nonresponse, coverage error, and processing error. Coverage error was not considered in

the article as the analysis was confined to differences between the interview modes for characteristics of the telephone population only. Any differential bias due to processing error was also eliminated as a major potential error source in the comparisons since care was taken during the data processing stage to apply essentially the same processing procedures to the data from both modes. Thus, differences in the estimates are attributable solely to measurement bias and nonresponse bias.

The article provided a study design and estimation methodology for estimating the mode biases associated with these two sources of error that do not rely on the existence of gold standard estimates or assumptions such as ''more is better.'' Rather, we used latent class analysis (LCA) to estimate the classification error probabilities for the telephone and face to face modes and obtained the measurement bias as a function of these estimates of the response probabilities. A key feature of the study measurement error evaluation design was the design of a telephone test-retest reinterview survey for both the CATI and face to face survey respondents.

To obtain estimates of the nonresponse bias, we relied on a telephone followup survey of the face to face survey nonrespondents. Using this information and the estimates of the response probabilities from the measurement error analysis, we showed that a consistent estimator of the true proportion, π, can be formed. This estimator of π was then used to estimate the bias due to nonresponse for both the face to face and CATI modes.

The components of mode bias estimation methodology were applied to data collected for the NHIS in Texas and California. Our analysis of the measurement error and nonresponse biases identified questions where the biases were significant even though the differences between prevalence estimates from the two modes were not significant. When the difference estimates were significant, our analysis provided information on the sources of the differences. Although the study was confined to only two states, the analysis provided evidence of substantial biases in both modes of interview for collecting health characteristics.

The analysis of mode biases suggests that neither CATI nor face to face interviewing is a uniformly superior mode of interview across all characteristics. For the characteristics considered in this analysis, measurement bias was considerable for both modes, often exceeding the nonresponse bias. The CATI nonresponse bias was often larger than that of face to face interviewing on average, but was often off-set by the measurement bias. This resulted in a smaller total bias for some characteristics. On the other hand, the measurement bias for the face to face mode was often larger than for the CATI mode. On average, the overall quality of the estimates from both modes was much the same.

These findings suggest that a CATI survey can produce data that compare well in quality to those produced by a face to face survey for many characteristics, despite a difference in response rates of more than 20 percentage points. There have been previous studies comparing NHIS face to face and telephone interviews with similar outcomes (see, for example, Cannell, Thornberry, and Fuchsberg 1981; Cannell et al. 1987). However, this study provides the first attempt to quantify the differences in both the measurement and nonresponse biases between the two modes of interview. Despite the limited scope of the study – only 14 survey questions were analyzed and the study was conducted in only two states – our results provide insight into the nature of the mode biases for

CATI and face to face and illustrate some important principles regarding the evaluation of these two important modes of interview.

There are ample opportunities for extending the current work. First, extending LCA to the other variables in the survey would provide additional evidence regarding the relative quality of face to face and CATI data. This analysis could also consider admitting additional exogenous variables into the LCA such as the respondent's age, race, sex, and other demographic characteristics. This would not only improve the fit of the models, but also provide additional information regarding the nature and causes of measurement error for both modes.

Second, there is still much to learn regarding the validity of the estimates from the LCA. Biemer (1997) provides some evidence of the validity of the LCA models for the analysis presented here. For example, as mentioned above, results from the test-retest reliability analysis described in the full report were consistent with the LCA results presented in Tables 6 and 7. The face to face survey data were of significantly lower reliability, on average, than the CATI data. Other work related to the validity of LCA for survey error evaluation can be found in Biemer (2000).

Third, to determine the generalizability of these results to the entire NHIS sample, a national study replicating the present study design is needed. Since our design allows both reinterview surveys and the nonresponse followup survey to be conducted by telephone, the design is relatively inexpensive to implement for an ongoing face to face survey.

Finally, we note that the survey research literature tends to judge the quality of a survey on the basis of the overall response rate. Our results suggest that equal emphasis should be given to the estimation and control of measurement errors in surveys. Quite often, a survey having a response rate of 80 percent may be perceived as having higher data quality than a survey of the same population with a response rate of only 60 percent. Our results suggest that such comparisons may be misguided and that factors such as the mode of interview, the design of the data collection process, and other features of the survey design that affect measurement error may be even more important than the final response rate in judging data quality.

6. References

Akaike, H. (1987). Factor Analysis and AIC. Psychometrika, 52, 317–332.

Aneshensel, C., Frerichs, R., Clark, V. and Yokopenic P. (1982). Telephone versus In-Person Surveys of Community Health Status. American Journal of Public Health, 72, 9.

Biemer, P.P. (1988). Measuring Data Quality. In Telephone Survey Methodology, eds. R. Groves, P. Biemer, L. Lyberg, J. Massey, W. Nicholls, and J. Waksberg, New York: John Wiley & Sons.

Biemer, P.P. (1997). Dual Frame NHIS/RDD Methodology and Field Test: Analysis Report. National Center for Health Statistics, Hyattsville, MD.

Biemer, P.P. (2000). On the Validity of Markov Latent Class Analysis for Estimating Classification Error in Labor Force Data. Survey Methodology, 26, 139–152.

Biemer, P.P. and Akin, D. (1994). The Efficiency of List-Assisted Random Digit Dialing Sampling Schemes for Single and Dual Frame Surveys. Proceedings of the Survey

Research Methods Section, American Statistical Association, Toronto, Canada, 1–10.

Biemer, P.P. and Stokes, S.L. (1992). Approaches to the Modeling of Measurement Error. In Measurement Errors in Surveys, eds. P.P. Biemer, R. Groves, L. Lyberg, N. Mathiowetz, and S. Sudman, 487–516. New York: John Wiley & Sons, Inc.

Bohrnstedt, G.W. (1983). Measurement. In Handbook of Survey Research, eds. P.H. Rossi, R.A. Wright, and A.B. Anderson, 70–122, New York: Academic Press.

Cannell, C., Thornberry, O., and Fuchsberg, R. (1981). Research on the Reduction of Response Error: The National Health Interview Survey. Silver Anniversary of the National Health Survey Act Contributed Papers, National Center for Health Statistics.

Cannell, C., Groves, R., Magilavy, L., Mathiowetz, N., and Miller, P. (1987). An Experimental Comparison of Telephone and Personal Health Interview Surveys. Vital and Health Statistics S2 106, Washington, DC: US GPO.

Chapman, D.W. (1976). A Survey of Nonresponse Imputation Procedures. Proceedings of the Section on Survey Research Methods, American Statistical Association.

Clogg, C.C. (1995). Latent Class Models. In Handbook of Statistical Modeling for the Social and Behavioral Sciences, eds. G. Arminger, C.C. Clogg, and M.E. Sobel, 311–359, New York: Plenum.

Cochran, W.G. (1977). Sampling Techniques. New York: John Wiley & Sons.

De Leeuw, E. and Van der Zouwen, J. (1988). Data Quality in Telephone and Face to Face Surveys: A Comprehensive Meta-Analysis. In Telephone Survey Methodology, eds. R.M. Groves, P. Biemer, L. Lyberg, J. Massey, W. Nicholls, and J. Waksberg, 283–300, New York: John Wiley & Sons.

Groves, R. (1989). Survey Costs and Survey Errors, John Wiley & Sons.

Groves, R. and Couper, M. (1998). Nonresponse in Household Interview Surveys. New York: John Wiley & Sons.

Groves, R. and Kahn, R. (1979). Surveys by Telephone: A National Comparison of Face-to-Face and Telephone Interviewing, New York: John Wiley & Sons.

Goodman, L. (1974). Exploratory Latent Structure Analysis Using Both Identifiable and Unidentifiable Models. Biometrika, 61, 215–231.

Hagenaars, J.A. (1988). Latent Structure Models with Direct Effects Between the Indicators: Local Dependence Models. Sociological Methods and Research, 16, 215–405.

Hansen, M.H., Hurwitz, W.N., and Pritzker, L. (1964). The Estimation and Interpretation of Gross Differences and the Simple Response Variance. Contributions to Statistics, ed. C.R. Rao, Calcutta: Pergamon Press Ltd., 111–136.

Hartley, H. (1962). Multiple Frame Surveys. Proceedings of the Social Statistics Section, American Statistical Association, 203–201.

Heinen, T. (1996). Latent Class and Discrete Latent Trait Models: Similarities and Differences. Thousand Oaks, California: Sage Publications.

Hidiroglou, M.A. and Gray, D.G. (1993). A Framework for Measuring and Reducing Nonresponse in Surveys. Survey Methodology, 19, 81–94.

Hui, S.L. and Walter, S.D. (1980). Estimating the Error Rates of Diagnostic Tests. Biometrics, 36, 167–171.

Körmendi, E. (1988). The Quality of Income Information in Telephone and Face-to-Face Surveys. In Telephone Survey Methodology, eds. R. Groves, P. Biemer, L. Lyberg, J. Massey, W. Nicholls, J. Waksberg, 341–357, New York: John Wiley & Sons.

Lin, T.H. and Dayton, C.M. (1997). Model Selection Information Criteria for Non-Nested Latent Class Models. Journal of Educational and Behavioral Sciences, 22, 3, 249–264.

Massey, J.T., Moore, T.F., Parsons, V.L., and Tadros, W. (1989). Design and Estimation for the National Health Interview Survey, 1985–1994. National Center for Health Statistics. Vital and Health Statistics, 2, 110.

Rao, J.N.K. and Thomas, D.R. (1991). Chi-Squared Tests with Complex Survey Data Subject to Misclassification Error. In Measurement Errors in Surveys, eds. P.P. Biemer, R. Groves, L. Lyberg, N. Mathiowetz, and S. Sudman, New York: John Wiley & Sons, 637–664.

Sinclair, M. and Gastwirth, J. (1996). On Procedures for Evaluating the Effectiveness of Reinterview Survey Methods: Application to Labor Force Data. Journal of the American Statistical Association, 91, 961–969.

Sirken, M.G. and Casady, R.J. (1987). Sampling Variance and Nonresponse Rates in Dual Frame, Mixed Mode Surveys. In Telephone Survey Methodology, eds. R. Groves et al., New York: John Wiley & Sons.

Sykes, W. and Collins, M. (1988). Effects of Mode of Interview: Experiments in the U.K. In Telephone Survey Methodology, eds. R. Groves, P. Biemer, L. Lyberg, J. Massey, W. Nicholls, and J. Waksberg, 301–320, New York: John Wiley & Sons.

Thornberry, O. and Massey, J. (1983). Coverage and Response in Random Digit Dialed National Surveys. Proceedings of the Section on Survey Research Methods, American Statistical Association, Toronto, Canada, 654–659.

Thornberry, O. and Massey, J. (1988). Trends in United States Telephone Coverage Across Time and Subgroups. In Telephone Survey Methodology, eds. R. Groves, P. Biemer, L. Lyberg, J. Massey, W. Nicholls, and J. Waksberg, 25–50, New York: John Wiley & Sons.

Trodahl, V.C. and Carter, R.E. Jr. (1964). Random Selecting of Respondents Within Households in Telephone Surveys. Journal of Marketing Research, 1, 71–76.

Turner, C.F., Ku, L., Rogers, S., Lindberg, L., Pleck, J., and Sonenstein, F. (1998). Adolescent Sexual Behavior, Drug Use, and Violence: Increased Reporting with Computer Survey Technology. Science, 280, 867–873.

U.S. Bureau of the Census (1985a). Evaluation of Censuses of Population and Housing, STD-ISP-TR-5, Washington, D.C., US GPO.

U.S. Bureau of the Census (1985b). Results of the 1984 NHIS/RDD Feasibility Study: Final Report. Internal Technical Report, submitted to the NCHS-Census Joint Steering Committee on Telephone Surveys, February.

U.S. Bureau of the Census (1994). Statistical Abstracts of the United States, 114th Edition, Washington, DC.

Van de Pol, F., Langeheine, R., De Jong, W. (1991). PANMARK User Manual, Version 2.2. The Netherlands Central Bureau of Statistics, Voorburg, The Netherlands.

Received February 2000
Revised January 2001