Mode Effect Analysis and Adjustment in a Split-Sample Mixed-Mode Web/CATI Survey

Stanislav Kolenikov, Courtney Kennedy
Abt SRBI, 55 Wheeler Street Cambridge, MA 02138 USA

Abstract

A major hazard in conducting multi-mode surveys is the potential for mode effects to compromise the response distributions recorded. In this study we evaluate the strengths and weaknesses of two approaches for statistically adjusting for mode effects. Under a regression modeling approach, adjustments are computed by regressing survey responses on mode, demographics, and other relevant variables. Under a multiple imputation approach, mode effects are conceptualized as a missing data problem. The imputation approach we pursue is based on an econometric framework of implied utilities in logistic regression modeling. We evaluate both approaches using data from the second wave of the Portraits of American Life Survey sponsored by Rice University's Kinder Institute for Urban Research. The mode of administration was randomly assigned as either CATI-only or Web with CATI follow-up for non-respondents. We detected a significant mode effect on four survey outcomes, after controlling for demographics and risk of type I error. The effects on the standard errors and point estimates are examined and discussed along with the advantages and disadvantages of each adjustment approach.

Key Words: mode effect, split-sample experiment, multiple imputation, logistic regression, social desirability

1. Introduction

As response rates for general population surveys continue to fall (Biener et al. 2004; Curtin et al. 2005; de Leeuw and de Heer 2002; Pew 2012), data collection agencies are looking for alternative ways to engage sample members and solicit response from them. One increasingly common strategy is to offer the sample units a variety of means to respond. Soliciting responses in the least expensive mode first and reserving more expensive modes for non-response follow-up can help minimize costs. Also, when designed properly, the provision of multiple modes can increase the overall survey response rate by making participation more convenient. From a Leverage-Salience Theory perspective (Groves et al. 2000), each time a new mode is introduced, the sample member's judgment about the relative benefits and drawbacks of participating can change. A related aspect is that some modes are more effective than others in achieving responses from certain groups (e.g., Web for reaching young adults).

One of the major drawbacks of multi-mode designs, however, is the potential for mode effects to undermine inference from the survey. A mode effects is a form of measurement error that occurs when respondents answer differently to a survey question solely because of the mode in which the question is administered (for example, Aquilino 1994; Hochstim 1967; Tourangeau and Smith 1996). Groves and his colleagues (2004) discuss various sources of mode effects including differences in frame coverage for each of the modes, auditive versus visual presentation of the questions, differences in handling of "don't know" or refusal response options, the role (if any) of the interviewer and consequences for social desirability pressure, and extremeness in choosing response options.

When it comes to estimation, survey designers typically deal with the potential for mode effects in one of two ways. In some instances, the issue is essentially ignored. For multi-mode surveys devoid of sensitive questions, questions with high item nonresponse rates, or other measurements prone to mode differences, the risk to estimates is arguably quite low. In other cases, survey designers attempt to measure mode effects by noting their direction and quantifying their magnitude at the question level. One rigorous technique for evaluating mode effects is to randomize mode assignment for at least a portion of the sample so that compositional differences by mode can be minimized (e.g., Heerwegh 2009; Kreuter et al. 2008). Under a randomized design, response distributions for each mode can typically be compared in straightforward analysis. Other approaches rely on statistical modeling. The goal of these approaches is to make the responding samples in each mode equivalent. This has been attempted by weighting (Lee 2006), using multivariate models (Dillman et al. 2009; Voogt and Saris 2005), and propensity score matching (Grodin and Sun 2008; Lugtig et al. 2011; de Vries et al. 2005).

2. Overview of Mode Effects Adjustments

While measuring mode effects is fairly common, actually adjusting survey estimates to correct for them is far less common. Our review identified only one federal or academic survey in the United States that incorporated a mode effect adjustment in the official survey estimates (Elliott et al. 2009). To be sure, in some instances the observed effects are too small to warrant adjustment. In other cases, though, survey researchers decline to adjust their data even in the face of significant mode effects for some estimates (Raglin et al. 2008; Soulakova et al. 2009; Quigley 2008).

The scant literature on performing mode effects adjustment (Christensen et al. 2006; Elliott et al. 2009; Powers et al. 2005) may have prevented some researchers from pursuing this activity. We also identified four other aspects of mode effects adjustment that likely explain why this is so rarely done in practice. From a scientific perspective, researchers should only implement an adjustment if there is a compelling reason to believe that the adjustment will make the survey estimate more accurate (less biased). If no “gold standard” validation data are available, it is difficult to determine whether the adjusted estimates are in fact more accurate than the unadjusted estimates. For questions about socially undesirable behaviors, which may suffer from underreporting, it may be reasonable to assert that higher estimates are more accurate. Similarly, socially desirable behaviors, such as volunteering or political involvement, may suffer from over-reporting, and so one might assert that lower estimates are more accurate.

Another potential barrier concerns variance. Any adjustment for mode effects has its own level of uncertainty. This uncertainty stems from the fact that the adjustments are based on survey samples that have sampling error. When a mode effects adjustment is incorporated into an estimate, the standard error of that estimate increases, as demonstrated in the analysis presented in this paper. Even if there is evidence that the adjustment makes the survey estimate more accurate, it may not necessarily reduce the mean square error of the estimate if the standard error is greatly inflated.

Two other drawbacks to mode effects adjustment concern costs and logistics. Testing for mode effects, developing adjustments, and computing the revised standard errors are complex tasks that require professional time from trained survey statisticians. Some survey instruments, including the one used in this paper, contain several hundred questions, which need to be tested, modeled, and analyzed individually. The resources required to perform that work entail substantial cost.

A related issue is that all of the statistical modeling and analysis takes time, lengthening the duration between the end of the field period and the release of the survey findings.

Some of the statistical work could perhaps be performed prior to the end of data collection, but some period of time for final analysis would inevitably be required.

In the next section, we discuss two different approaches to this problem. This is followed by a description of the survey to which both approaches were applied and then an analysis comparing the approaches. Several conclusions from the analysis are discussed.

3. Approaches for Mode Effects Adjustment

We start with a straw man example to motivate the more complex adjustments evaluated in this analysis. Perhaps the simplest mode adjustment that can be thought of would consist of computing the mean response in each mode, quantify the difference between the less accurate mode(s) and the reference mode, and then subtracting that difference from the responses in the less accurate mode(s). Such an approach would only be valid under a number of restrictive assumptions. The variable of interest must be continuous. Also, there should be no differential non-response between the different survey modes. In most practical applications, however, categorical variables will be encountered, and it would be inappropriate to subtract a fractional quantity from a 0/1 response. Finally, different demographic groups may have different propensities to respond in each mode (e.g., the younger respondents may prefer to use web, and the older ones, CATI). Thus, a robust mode adjustment must be able to deal with an arbitrary response scale, and control for potentially different coverage or different response propensities for different modes.

In this study we evaluate the strengths and weaknesses of two approaches for statistically adjusting for mode effects. Under a regression modeling approach, adjustments are computed by regressing survey responses on mode and respondent demographics. Under a multiple imputation approach, mode effects are conceptualized as a missing data problem. The imputation approach we pursue is based on an econometric framework of implied utilities in logistic regression modeling. Both approaches assume that the responses collected in one mode are more accurate than the responses in the other modes(s). The more accurate mode is treated as the “benchmark” on which the adjustments are based.

3.1. Regression adjustment

The regression adjustment approach consists of fitting a regression model that includes the mode(s) as predictor(s), along with additional relevant variables, and harmonizing the responses across modes by subtracting the estimated coefficient of a given mode from the (aggregated) response. In this context relevant variables are those that covary with both mode and the survey variable of interest.

To formalize this argument, consider a regression model

$$\mathrm{y_i} = \beta' \mathrm{x_i} + \gamma \mathrm{m_i} + \varepsilon_\mathrm{i} \tag{1}$$

where y is a continuous response of interest, x are demographic predictors, m is the survey mode indicator (cases in the “benchmark” mode are assigned 0 for this indicator), β is the vector of regression coefficients, γ is the mode effect, and ε is the regression residual. Suppose that the model (1) is fitted by ordinary least squares (OLS) regression or by weighted least squares using survey weights, and coefficient estimates $\hat{\beta}$ and $\hat{\gamma}$ are obtained. Then the regression adjustment is the predicted value with the mode contribution $\hat{\gamma} m$ excluded:

$$\tilde{\mathrm{y}}_\mathrm{i} = \mathrm{x}_\mathrm{i}' \hat{\beta} + \hat{\varepsilon_\mathrm{i}} = \mathrm{y_i} - \hat{\gamma} \mathrm{m_i} \tag{2}$$

where $\hat{\varepsilon}$ is the regression residual. Such an adjustment treats the mode effect as known, rather than as estimated, requires corrections for standard errors, and is only applicable to continuous variables. An example of using the regression approach to adjust for mode effects is the work performed by Elliott and his colleagues (2009) for the Consumer

Assessments of Healthcare Providers and Systems Hospital Survey (Hospital CAHPS®), in which hospital performance metrics were harmonized across four modes: mail, telephone, mail with telephone follow-up, and interactive voice response (IVR), accounting for the differences in the composition of patients responding in each mode (referred to as “patient mix” by Elliott et. al. (2009).

While the regression approach works well for continuous and, in some cases, ordinal variables, it does not easily generalize to categorical data. An extension of the regression mode adjustment to logistic regression is possible only for the summaries of the data, such as estimated proportions. As is known from the general theory of logistic regression modeling (Maddala 1986, Sec. 2.5), the sum of predicted probabilities is equal to the sum of outcomes, which also carries over to the weighted logistic regression model. Thus, if the 0/1 response y_i is modeled as

$$\mathrm{Prob}[\, \mathrm{y_i} = 1 | \mathrm{x_i}, \mathrm{m_i}] = \Lambda(\beta' \mathrm{x_i} + \gamma \mathrm{m_i}), \Lambda(\mathrm{u}) = (1 + \exp(\, -\mathrm{u}))^{-1} \qquad (3)$$

and predicted probabilities are formed as

$$\tilde{p}_i = \Lambda\big(\hat{\beta}' x_i\big) \qquad (4)$$

then the sum of the predicted probabilities can be viewed as the mode-adjusted incidence:

$$\tilde{p} = \textstyle\sum_i \tilde{p}_i = \sum_i \Lambda\big(\hat{\beta}' x_i\big) \qquad (5)$$

The standard errors for this $\tilde{p}$ can be obtained using the delta method. The point estimates and the standard errors it reports are for the average probability of the “Yes” response, which may only be meaningful for the estimate for the full sample. The standard errors are model-based, rather than design-based (Binder and Roberts 2003, 2009). Unlike the mode adjustment (2) for continuous variables, adjustment (5) does not involve the actual responses. If the model does not contain important covariates or otherwise lacks predictive power, the estimates may be biased towards the overall mean. This pattern is observed and discussed in the Results section.

As the probability of an event can be thought of the expected value of the corresponding 0/1 Bernoulli variable, the linear regression mode adjustment (2) can also be nominally applied to the binary response data using the linear probability model. While this model is known to suffer from a number of drawbacks, such as predictions out of natural ranges, nonlinearity and heteroskedasticity (Maddala 1986, Long 1999, Wooldridge 2010), some authors argued in favor of its use in certain situations (Angrist and Pischke 2008).

3.2. Multiple imputation adjustment

A different approach to mode effects adjustment conceptualizes it as a missing data. The imputation approach we pursue is based on an econometric framework of implied utilities in logistic regression modeling, and features multiple imputation (Rubin 1996, 2004) for the mode effect adjustment and the accompanying variance estimation. This methodology has a clear appeal for mode effect adjustments. Powers and colleagues (2005) used multiple imputation in its classical form to adjust for differences in reported health using the SF-36 scale between mail and telephone modes of Australian Longitudinal Study of Women’s Health. Christensen and colleagues (2006) used ideas of test equating from psychometric item-response theory to adjust sum scores for a psychometric scale, i.e., several dichotomous items considered together, usually by a simple sum. Peytchev (2012) used multiple imputation to improve estimates on sensitive question on abortions in National Survey of Family Growth, based on a rich frame data from NHIS. The mode adjustment he performed was to modify responses in the more sensitive CAPI mode compared to the responses in self-administered ACASI mode. Of these, Christensen et. al. (2006) used a split-sample design with random assignment of respondents to mode, while Powers et. al. (2005) and Peytchev (2012) were observational studies that relied on self-selection into interviewing mode.

While multiple imputation is a promising procedure for many applications and has been used in some large scale government surveys (Barnard and Meng 1999), the method has limitations when applied with complex survey data (Fay 1996; Kim et. al. 2006; Reiter et. al. 2006). First, the survey data often contain complicated violations of the i.i.d. assumption that multiple imputation has to invoke, at least implicitly. Second, multiple imputation supports model-based inference, while in finite population surveys the interest is generally in design-based inference. Generally, to apply multiple imputation to mode adjustment, one must discard the existing data on the affected variables in the less accurate mode. In the next section we propose an improved procedure that retains some of the information contained in the observed response.

3.3. Implied utility – multiple imputation mode adjustment

To develop the new adjustment, let us invoke the latent variable approach to limited dependent variable modeling typically used to introduce logistic and ordinal logistic regression models in economics and some other social sciences (Maddala 1986; Long 1997; Wooldridge 2010). For a logistic regression model, the observed 0/1 outcome y_i is viewed as a crude reflection of an underlying propensity y_i^* to endorse a positive response:

$$\mathrm{y_i} = \begin{cases} 1, \mathrm{y_i^*} > 0 \\ 0, \mathrm{y_i^*} \leq 0 \end{cases} \qquad (6)$$
$$\mathrm{y_i^*} = \beta' \mathrm{x_i} + \gamma \mathrm{m_i} + \varepsilon_\mathrm{i}, \varepsilon_\mathrm{i} \sim \Lambda(\cdot)$$

where the error term ε_i follows a logistic distribution with cdf $\Lambda(z) = 1/[1 + \exp(-\mathrm{z})]$. The underlying latent variable y_i^* is interpreted as the utility associated with choosing a positive response, $y_i = 1$, and is compared to the “reservation utility” of zero associated with the negative response, $y_i = 0$. This model is equivalent to (3), but has a different motivation, which happens to be more useful in the context of mode effects. Unlike in the linear regression model where the residual can be computed explicitly and accurately as $\widehat{\varepsilon_i} = y_i - \hat{y}_i$, the information provided by say the response $y_i = 1$ is only that $y_i^* > 0$, i.e., that $\varepsilon_i > -\beta' x_i - \gamma m_i$. Unlike the case of the continuous data (2), we cannot say with certainty what the observed value $\tilde{y}_i$ is going to be after the mode adjustment, since the underlying value of y_i^* is unknown. However, we can exploit the information that, conditional on the observed response, the regression residual follows a truncated logistic distribution. Namely, we can draw the residuals from this distribution, apply them to the fixed effect part to obtain the simulated y_i^*, remove the estimated mode effect $\hat{\gamma} m_i$, and compare the result to zero to come up with the mode-adjusted value $\tilde{y}_i$. Since a single simulation will be subject to simulation noise, the procedure can be repeated multiple times to obtain plausible values of $\tilde{y}_i$. In effect, this is a multiple imputation procedure in which a highly specialized model is developed for the imperfectly observed data y_i. Unlike the more mainstream multiple imputation procedures for binary data that would ignore the observed data y_i and simulate $\tilde{y}_i \sim \text{Bernoulli}(\tilde{p}_i)$ where $\tilde{p}_i$ is the mode effect adjusted probability of the positive response (4), the proposed procedure retains additional information regarding whether the residual ε_i is likely to be high or low, based on the observed response.

An extension of this implied utility model to an ordinal response treats a finer mesh of ranges into which the utility y_i^* could fall:

$$\mathrm{y_i} = \begin{cases} 1, -\infty = \tau_0 < \mathrm{y_i^*} \leq \tau_1 \\ 2, \quad \tau_1 < \mathrm{y_i^*} \leq \tau_2 \\ \quad \cdots \\ \mathrm{K}, \quad \tau_{\mathrm{K}-1} < \mathrm{y_i^*} < \tau_{\mathrm{K}} = +\infty \end{cases} \tag{8}$$

$$\mathrm{y_i^*} = \beta'\mathrm{x_i} + \gamma\mathrm{m_i} + \varepsilon_\mathrm{i}, \varepsilon_\mathrm{i} \sim \Lambda(\cdot)$$

Note that the set of regressors does not contain the intercept, otherwise the thresholds $\tau_1, \dots, \tau_{K-1}$ will not be identified. Like in the ordinary logistic model, the observed response y_i provides limited information about the location of the implied utility, and residuals conditional on the observed response can be drawn from truncated logistic distribution for multiple imputation purposes. Thus, to provide mode corrections for ordinal Likert scale variables, the following algorithm can be used:

3.4. IUMI Algorithm.

1. Estimate the ordinal logistic regression model with the response of interest y_i suspect to suffer from the mode effect using model (8) with the appropriate weights, and obtain the parameter estimates $\hat{\beta}$ and $\hat{\gamma}$.
2. For *m*-th imputation, *m*=1, …, *M*, simulate the residual from the truncated logistic distribution:

$$\hat{\varepsilon}_\mathrm{i}^{(m)} = \Lambda^{-1}\{\tau_{k-1} + (\tau_k - \tau_{k-1})U\} - \left(x_i'\hat{\beta} + \hat{\gamma}m_i\right), \mathrm{y_i} = \mathrm{k}, \ U \sim U[0,1] \tag{9}$$

3. Form implied utility that has the mode effect removed:

$$y_i^{*(m)} = \hat{\beta}'x_i + \hat{\varepsilon}_\mathrm{i}^{(m)}$$

4. Form the imputed value of the response:

$$\mathrm{y_i}^{(m)} = \begin{cases} 1, -\infty = \tau_0 < y_i^{*(m)} \leq \tau_1 \\ 2, \quad \tau_1 < y_i^{*(m)} \leq \tau_2 \\ \quad \cdots \\ K, \quad \tau_{K-1} < y_i^{*(m)} < \tau_K = +\infty \end{cases}$$

5. Conduct the analysis of interest, such as tabulation, and store the point estimates and variance-covariance matrices.
6. Repeat steps 2–5 a sufficiently large number of times *M*.
7. Combine the results stored at step 5 using Rubin’s multiple imputation rules.

The regression model in step 1 can be estimated on a subset of cases if it is more appropriate within the context of the study design. The subsequent steps can be applied to all observations in the data, if needed. Logistic regression is a special case with just one threshold parameter τ_1.

An extension of the implied utility framework is also possible for the multinomial logistic regression model. One category is taken as a baseline with zero coefficients, and other categories are compared to it in a manner similar to (7). However, the distribution of simulated residuals becomes complicated by conditioning on the chosen category. For example, if the fixed part of the implied utility is the same for all categories, and is equal to zero, then the chosen category must have received a larger residual, while other categories must have residuals smaller than the one in the chosen category. The appropriate conditional distributions were given by Train and Wilson (2008).

4. Research Design

We compared the regression and IUMI approaches to mode effects adjustment using the second wave of the Portraits of American Life Study, a panel survey conducted by Abt SRBI for the Kinder Institute for Urban Research at Rice University. The survey measured religious identification, congressional affiliation and participation, as well as religious and political beliefs. The average interview length was 75 minutes. The first wave was conducted in 2006 with a national area probability sample with oversampling of ethnic and racial minorities (African Americans, Hispanics, and Asians). That wave featured interviews with 2,610 respondents age 18 or older. In the second wave, conducted from March 12 to September 30, 2012, n=1,320 respondents were successfully re-interviewed. In additional, 389 young adults who were children under age 18 in 2006 had become eligible for the interview in 2012, and 101 of them were interviewed. For more information and to download the publicly available and restricted data sets, visit the study website http://www.thearda.com/pals/.

The primary data collection mode for the second wave was self-administered Web survey, and CATI was used to follow up and sample members not responding on the Web. Given the potentially sensitive nature of certain questions and the fact that both self-administered and interviewer-administered modes were planned, the risk of mode effects was carefully examined at the design stage. The potential threat from mode effects was addressed by assigning a randomly selected fraction of the wave 2 sample (13%) to a CATI-only condition that did not feature a Web response option. This was done because a mode effects analysis that simply compared the Web completions to the completions from the CATI follow-up would have been confounded by the fact that the CATI completions were less amenable to taking the survey and may have systematically differed on variables in the survey. The most rigorous way to avoid such confounding was to assign a randomly selected portion of the sample to complete the survey by CATI. This experimental design yields an identifiable set of respondents in each mode who completed the survey in that mode by virtue of random assignment.

Toward the end of the field period, the decision was made to allow the sample units randomized into CATI-only condition to complete the survey on the Web. This measure was taken to increase the response rate. Among the completions in the CATI-only condition, 93 responded by CATI and 72 responded by Web. The 72 CATI-to-Web completions were excluded from this mode effects analysis, so as to avoid contaminating the randomized mode comparison.

The panel nature of the sample alleviates the issues of the differences in frames and coverage. Hence, most of the remaining mode effects would be due to access to the different modes, and measurement properties of the different formats of the presentation (visual on the computer screen versus sequential on the phone) and presence of the interviewer (in the CATI mode). Table 1 reports the number of cases assigned to each mode condition, the number of completions, and the response rate.

Table 1. Sample Size, Completions, and Response Rate by Mode

Mode Condition	Cases Assigned to Mode Condition	Completions	Response Rate (AAPOR(1))
Web with CATI follow-up	2,934		42.8%
Completed by Web		1,102	
Completed by CATI		154	
CATI only*	391		42.2%
Completed by CATI		93	
Completed by Web (excluded from analysis)		72	
Not contacted	54		
Total	3,325	1,421	42.7%

*Note: In the end of the field period, the nonresponding cases in the CATI-only condition were invited to complete the survey on the Web. These cases were excluded from the mode effects analysis but are shown here for a full description of the survey.

For the mode effects analysis, the respondents of interest are those completing by Web in the Web with CATI follow-up condition (n=1,102) and those completing by CATI in the CATI only condition (n=93). The number of completions in the latter group is relatively small, which indicates that the size of the standard errors in the mode comparison and adjustments will be a concern. In other words, only fairly strong mode effects will likely be detected. Both the regression and the multiple imputation methods are affected in roughly the same degree by sampling variability of the mode effect estimate, $\hat{\gamma}m_i$, as they both rely on taking this quantity out of their respective equations in order to remove the estimated effect from mode.

The overriding concern motivating the analysis of mode effects in wave 2 was greater social desirability pressure in the CATI interviews leading to some editing of responses.[1] For example, the mode effects literature suggests that reporting of activities such as religious attendance and political engagement might be higher in the CATI mode than in the Web mode due to the presence of an interviewer and associated social desirability pressure (e.g., Kreuter et al. 2008; Presser and Stinson 1998). The assumption applied to the mode effects analysis was, thus, that the Web responses were, if anything, likely to be more accurate than the CATI responses.

5. Results

5.1. Testing for Mode Differences

As the first step in determining the impact of mode on the survey results, the substantive variables were first cross-tabulated against the mode of interview for the CATI-only and Web completions for the subsample of compliers, i.e., the respondents who completed the survey in the mode originally assigned to them. Rao and Scott (1981) corrected *p*-values were computed, and a false discovery rate multiple testing procedure (Benjamini and Hochberg 1995) at a relatively liberal .10 significance level was used to determine which variables exhibited statistically significant marginal differences between modes, without controlling for demographics or other factors. The Benjamini-Hochberg procedure rejects the null hypotheses for which $P_{(k)} \leq \frac{k}{m}\alpha$, where m is the total number of hypotheses to be tested, α is the target significance level, and the p-values are ordered in increasing order, $0 \leq P_{(1)} \leq P_{(2)} \leq \cdots \leq P_{(m)} \leq 1$. Out of the full list of 297 behavioral and attitudinal variables, 278 variables did not show detectable differences between the marginal distributions in the two modes; and 19 variables demonstrated statistically significant differences between modes (after controlling for type I error in multiple testing). Thus the largest *p*-value for which a hypotheses was not rejected was $19 \cdot 0.10\ /\ 297\ = 0.0064$ at the overall 10% significance level. Of these, 3 variables had fewer than half complete cases, and were removed from subsequent analysis.

The remaining 16 variables are listed in Table 2, in the order of significance indicated by p-values of Rao-Scott test. The agree/disagree 5-point Likert scale had options from “Strongly disagree” options first to “Neither agree nor disagree” in the middle to

[1] There was also concern about mode effects compromising longitudinal analysis because CAPI and self-administered paper questionnaires (SAQ) were used in wave 1 (for different modules), but wave 2 was conducted using Web and CATI. As it was impossible to replicate the wave 1 modes in wave 2, so isolating the effect from wave 1 mode would have been extremely difficult. The potential for mode effects in the longitudinal analysis is, thus, beyond the scope of this paper.

"Strongly agree". The order of options was different in different questions. The satisfaction 5-point scale was presented from "Very satisfied" to "Very dissatisfied" response categories, with the neutral category in the middle. The order of categories varied between groups of questions: some had "Strongly agree" upfront, and some had "Strongly disagree" upfront, as indicated in the table. At face value, most of these questions (except for hair color) have a notable social desirability component and/or relate to the current social issues that are actively debated in the society.

Some questions had a non-monotonic pattern of differences between the modes that cannot be adequately modeled in either the regression or IUMI approach. For instance, the questions on supernatural miracles and divine healing have more extreme responses in CATI mode. The question on abortion had 5% refusals in CATI (as compared to no refusals on the Web), and respondents could volunteer their description of the "case-by-case" judgment. The Web instrument, on the other hand, explicitly had "Other" text box that was used by 2.5% of respondents.

These 16 survey items were then subjected to multivariate analysis to determine whether the mode effect could be explained by differences in the demographic distributions of the respondents in the two key mode conditions. For example, 44% of the Web respondents were under age 40, compared to just 22% of the CATI only respondents (a significant difference at the .02 level). An appropriate (logistic, ordinal logistic, or linear) regression model was fit based on the Web respondents and CATI-only respondents. Besides the survey mode, the explanatory variables included gender, marital status (4 categories), presence of children in the household, employment status, race/ethnicity (4 categories), and a quadratic function of age. The p-value of the mode coefficient in the regression model is reported in the last column of Table 2. The false discovery rate multiple testing adjustment was again repeated with 10% overall cutoff and four items shown with asterisks in Table 2 have demonstrated significance according to this criterion.

5.2. Regression and implied utility-multiple imputation adjustment

The regression and implied utility–multiple imputation (IUMI) adjustments were tested with the four items showing a significant mode effect even after controlling for the demographic variables in the regression model. The number of close friends question was treated as a continuous variable. The mode effect estimate is 1.17 more friends reported in CATI (s.e. = 0.50). The other three questions are binary variables, and mode effect adjustments for them were calculated using the regression mode adjustment and IUMI adjustment using Algorithm 1. The number of imputations taken was M=20 to ensure that the minimal degrees of freedom (Barnard and Rubin 1999; Reiter 2007) is comparable to or exceeds the nominal sample size.

The results are reported in Table 3. For each variable, we conducted the analysis of the original data using the appropriate survey design with weighting, stratification, and clustering, and the linear regression adjustment (2). For the binary items, the logistic regression adjustment (5) and the IUMI adjustment were also applied. Recall that the regression adjustments (5) for the binary variables are weighted sums of predicted probabilities, rather than the weighted average responses. They use the information from the pooled sample of Web respondents and CATI-only respondents, with n=1,151, and thus their standard errors are comparable to that of the overall direct estimate.

We also applied the linear regression adjustment (2) by formulating a linear probability model, estimating it with weights, and subtracting the estimated mode effect from the direct estimates. This is the form of correction applied by Elliott et al. (2009). The reported standard errors were corrected for sampling variability by performing survey bootstrap (Rao, Wu and Yue 1992; Kolenikov 2010) on the adjusted estimate based on (2).

Table 2. Variables with potential mode effects

Survey item	Nature of the mode effect	n	Rao-Scott p-value	Regres-sion p-value
What color is your hair today?	More "Grey/white" and "Other" (open ended) in CATI. More "Brown" and "Blonde" on the Web.	928	.000044	N/A
Do you personally believe that abortion should be legal under...	More "Some" circumstances, more refusals in CATI. More "Almost all", "Most" and "No" circumstances, more "Other" on the Web	1182	.000341	0.7354
Year of birth	Older in CATI	1151	.000546	N/A
In the past 12 months have you helped directly by giving some of your time to close family?	More "Yes" in CATI	1184	.000765	0.0018*
In the past 12 months have you helped directly by giving some of your time to neighbors?	More "Yes" in CATI	1184	.000843	0.0094*
One of the most effective ways to improve race relations in the U.S. is to stop talking about race.	More "Somewhat disagree" and "Strongly agree" (last category) in CATI	1178	.001302	0.6588
An angel has directly helped me in a time of need.	More "Somewhat" or "Strongly agree" (last) in CATI	1176	.001488	0.3117
Immigrants coming into the US are taking too many jobs away from other American citizens.	More "Somewhat disagree" and "Somewhat" or "Strongly agree" (last) in CATI	1179	.001509	0.7089
How satisfied or dissatisfied are/were you with sermons, preaching, or homilies at your congregation?	More "Somewhat dissatisfied" and "Not applicable" (last) in CATI	783	.001728	0.0646
How much respect do you have for a head pastor / the religious leadership?	More "A little bit" and "None at all" (last) in CATI	783	.001971	0.8531
I have experienced a supernatural miracle, an event that could not have happened without the intervention of God or a spiritual force.	More "Strongly agree" (first), "Somewhat disagree" and "Strongly disagree" (last) in CATI	1179	.002413	0.3214

In the past 5 years, have you had a major financial crisis?	More "No" in CATI	929	.003061	0.0120*
I have experienced or witnessed a divine healing of an illness or injury.	More "Strongly agree" (first), "Somewhat disagree" and "Strongly disagree" (last) in CATI	1176	.003170	0.4235
Not including people living in your home, about how many people, if any, would you say you feel close to?	More 6+ in CATI	1187	.004187	0.0218*
I believe in reincarnation, that people have lived previous lives.	More "Somewhat disagree" and "Strongly disagree" (last) in CATI	1178	.004689	0.0646
How satisfied or dissatisfied are you with religious education classes for adults, such as Sunday, Church, or Sabbath school, Bible class, Quran class, etc.?	More "Not applicable", "Very satisfied" (first) or "Somewhat satisfied" in CATI	785	.004894	0.6752

Note: * significant at 10% level controlling for the false discovery rate.

The implied utility-multiple imputation correction produced estimates that were more similar to the linearly adjusted estimates than to the estimates adjusting using (5). The standard errors on the adjusted estimate, as well as the overall estimate, are higher than on the original data. We believe this is more plausible than a sharp decline in the standard errors reported by Peytchev (2012). The additional increase in estimated variability is due to the sampling error of the mode adjustment introduced into the data. An adjustment that accounts for the standard error of $\hat{\gamma}$ was made, where steps 2–3 of the IUMI algorithm draw from the distribution $N(\hat{\gamma}, [\mathrm{s.e.}\,\hat{\gamma}]^2)$ instead of using a fixed value $\hat{\gamma}$. Our experimental use of this adjustment showed that its effect is in the third decimal point in the standard error.

Overall, the IUMI adjustment appeared to perform best, producing the shifts in the expected direction, and an expected increase in the standard errors. The synthetic estimate based on (5) may suffer from bias of the adjusted estimates towards the mean. When applied to the binary data, the linear regression adjustment (2) produced estimates comparable to those of IUMI, but the standard errors were larger for most statistics, including the overall one.

6. Discussion

This paper evaluated two approaches for statistically adjusting for mode effects in a mixed mode survey. The initial analysis of the mode effects was greatly simplified by the random assignments of the sample units into different modes. Otherwise, an additional model for propensity to respond in a given mode would have to have been fitted to the data, and/or a Heckman-type model (Wooldridge 2010) be used for item(s) of interest.

The advantage of the proposed implied utility–multiple imputation adjustment is internal consistency under no mode effect. We have observed that the logistic regression adjustment modified all the modes, and probably shrunk the estimates towards the grand mean too much. The IUMI adjustment, however, has the desirable property of not altering the responses in the benchmark mode. Moreover, when no mode effect is present, the IUMI adjustment maintains the original data, while a typical multiple imputation procedure would discard the original data and simulate the response using the predicted probability, thus increasing the simulation noise in the multiply imputed data and inflating the between-simulation variance.

Table 3. Mode effect adjustments

	CATI only	Web only	CATI -> Web	Web -> CATI	Overall
In the past 12 months have you helped directly by giving some of your time to close family? (% Yes)					
Without adjustments	92.6 (2.7)	76.2 (1.9)	73.2 (7.1)	75.3 (4.4)	77.1 (1.6)
Logistic regression adjustment (5)	75.4 (2.0)	76.3 (1.9)	77.2 (1.7)	77.2 (2.3)	76.4 (1.8)
Linear regression adjustment (2)	75.7 (2.4)	76.2 (1.9)	73.2 (7.2)	58.4 (6.0)	74.4 (1.9)
Implied utility – MI adjustment (Algorithm 1)	73.0 (9.1)	76.2 (1.9)	73.2 (7.1)	60.2 (6.4)	74.4 (1.8)
In the past 12 months have you helped directly by giving some of your time to neighbors? (% Yes)					
Without adjustments	59.8 (6.4)	36.0 (2.5)	36.4 (7.6)	49.0 (5.7)	38.8 (2.0)
Logistic regression adjustment (5)	38.7 (2.8)	36.3 (2.4)	37.1 (2.3)	37.5 (3.2)	36.6 (2.3)
Linear regression adjustment (2)	37.2 (4.0)	36.0 (2.4)	36.4 (7.6)	28.8 (8.3)	35.4 (2.5)
Implied utility – MI adjustment (Algorithm 1)	38.0 (9.5)	36.0 (2.5)	36.4 (7.6)	31.4 (6.5)	35.7 (2.1)
In the past 5 years, have you had a major financial crisis? (% Yes)					
Without adjustments	14.2 (4.6)	34.9 (2.7)	45.3 (9.2)	21.3 (5.2)	32.9 (2.3)
Logistic regression adjustment (5)	32.1 (2.8)	35.5 (2.6)	40.1 (3.1)	29.6 (2.7)	35.0 (2.5)
Linear regression adjustment (2)	30.7 (4.2)	34.9 (2.6)	45.3 (9.0)	37.8 (9.0)	35.3 (2.7)
Implied utility – MI adjustment (Algorithm 1)	30.6 (9.6)	34.9 (2.7)	45.3 (9.2)	35.4 (7.4)	35.2 (2.5)
Number of persons outside your home that you feel closest to (average)					
Without adjustments	8.2 (0.6)	6.5 (0.2)	7.7 (0.5)	7.2 (0.5)	6.8 (0.2)
Linear regression adjustment (2)	6.9 (0.3)	6.5 (0.2)	7.7 (0.5)	6.0 (0.7)	6.6 (0.2)

Any of the proposed adjustments imply additional analytical and/or statistical programming work. The linear regression adjustment is relatively straightforward, but would require replicate variance estimation. Procedurally, the adjustment routine must be isolated into a separate piece of code that would take the response variable, the mode variable, the demographic controls, and the replicate weights as inputs, and produce estimated proportions of interest as the output, to be combined by the standard or custom code for replicate variance estimates. The IUMI adjustment may be used with the standard MI routines, but it requires custom programming of the imputation procedure.

These programming steps can be easily performed in environments suitable for custom programming, such as Stata or R, in which either survey-based estimation or multiple imputation estimation can be applied to an arbitrary estimation procedures. Implementing the proposed approaches in SAS or SPSS that lack replicate weights estimation and/or flexibility to create your own imputed data would require programming everything from scratch.

Acknowledgments

The Portraits of American Life Survey was sponsored by Rice University's Kinder Institute for Urban Research. We would like to thank Mike Emerson, Adele James, Chintan Turakhia and Ali Ackermann for their leadership on Portraits of American Life Survey and helpful comments that we received from them.

References

Angrist, J. D. and J.-S. Pischke (2008). Mostly Harmless Econometrics: An Empiricist's Companion. Princeton University Press.

Aquilino, W. (1994). Interview Mode Effects in Surveys of Drug and Alcohol Use. Public Opinion Quarterly, 57: 358-376.

Barnard, J. and X.-L. Meng (1999). Applications of multiple imputation in medical studies: from AIDS to NHANES. *Statistical Methods in Medical Research 8* (1), 17-36.

Barnard, J., and D. B. Rubin. (1999). Small-sample degrees of freedom with multiple imputation. Biometrika 86, 948–955.

Biener, L., Garrett, C.A., Gilpin, E.A., Roman, A.M., and Currivan, D.B. (2004). Consequences of Declining Survey Response Rates for Smoking Prevalence Estimates. American Journal of Preventative Medicine, 27(3), 254-257.

Benjamini, Y. and Y. Hochberg (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological) 57* (1), 289-300.

Binder, D. A. and G. R. Roberts (2003). Design-based and model-based methods for estimating model parameters. In R. L. Chambers and C. J. Skinner (Eds.), *Analysis of Survey Data*, Chapter 3. New York: John Wiley & Sons.

Binder, D. A. and G. Roberts (2009). *Design- and Model-Based Inference for Model Parameters*, in: D. Pfeffermann and C. R. Rao (eds), Handbook of Statistics, Volume 29B, Sample Surveys: Inference and Analysis, pp. 33-54. Elsevier.

Christensen, K.B., Feveile, H., Kreiner, S. and Bjorner, J.B. (2006) Adjusting for mode of administration effect in surveys using mailed questionnaire and telephone interview data. Research Report, Department of Biostatistics, University of Copenhagen, Denmark.

Curtin, R., Presser, S. and Singer, E. (2005). Changes in telephone survey nonresponse over the past quarter century. Public Opinion Quarterly, 69, 87–98.

de Leeuw, E., and de Heer, W. (2002), "Trends in Household Survey Nonresponse: A Longitudinal and International Comparison," in Survey Nonresponse, eds. R. M. Groves, D. A. Dillman, J. L. Eltinge, and R. J. A. Little, New York: John Wiley & Sons, pp. 41–54.

de Vries, H., Elliott, M.N., Hepner, K.A., Keller, S.D., and Hays, R.D. 2005. "Equivalence of Mail and Telephone Responses to the CAHPS® Hospital Survey. Health Services Research, 40: 2120-2139.

Dillman, D.A., Phelps, G., Tortora, R., Swift, K., Kohrell, J., Berck, J. & Messer, B.L. (2009) Response Rate and Measurement Differences in Mixed-Mode Surveys Using

Mail, Telephone, Interactive Voice Response (IVR) and the Internet. Social Science Research 38, 1, pp. 1–18.

Elliott, M. N., Zaslavsky, A. M., Goldstein, E., Lehrman, W., Hambarsoomians, K., Beckett, M. K. and Giordano, L. (2009), Effects of Survey Mode, Patient Mix, and Nonresponse on CAHPS Hospital Survey Scores. Health Services Research, 44: 501–518. doi: 10.1111/j.1475-6773.2008.00914.x

Fay, R. E. (1996). Alternative paradigms for the analysis of imputed survey data. *Journal of the American Statistical Association 91* (434), 490-498.

Groves, R. M., Fowler, F.J., Couper, M.P., Lepkowski, J.M., Singer, E., and Tourangeau, R. (2004). Survey Methodology. Wiley Series in Survey Methodology. New York: John Wiley and Sons.

Groves, R.M., Singer, E., and Corning, A. (2000). Leverage-Salience Theory of Survey Participation. Public Opinion Quarterly 64: 299-308.

Heerwegh, D. (2009). Mode Differences Between Face-to-Face and Web Surveys: An Experimental Investigation of Data Quality and Social Desirability Effects. International Journal of Public Opinion Research, 21: 111-121.

Hochstim, J. (1967). A Critical Comparison of Three Startegies of Collecting Data from Households. Journal of the American Statistical Association, 62: 976-989.

Kim, J. K., Brick, M. J., W. A. Fuller, and G. Kalton. (2006). On the bias of the multiple-imputation variance estimator in survey sampling. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 68* (3), 509-521.

Grodin, C. and Sun, L. 2008. 2006 Census Internet Mode Effect Study. Proceedings of the Section on Survey Research Methods Section, American Statistical Association.

Kolenikov, S. (2010). Resampling variance estimation for complex survey data. The Stata Journal 10 (2), 165-199.

Kreuter, F., Presser, S., and Tourangeau, R. (2008). Social Desirability Bias in CATI, IVR, and Web Surveys: The Effects of Mode and Question Sensitivity. Public Opinion Quarterly 72: 847-865.

Lee, S. (2006). Propensity Score Adjustment As a Weighting Scheme for Volunteer Panel Web Surveys. Journal of Official Statistics, 22: 329-349.

Long, J. S. (1997) Regression Models for Categorical and Limited Dependent Variables. Advanced Quantitative Techniques in the Social Sciences Series, SAGE, Thousand Oaks, CA.

Lugtig, P.J., Lensvelt-Mulders, G.J.L.M, Frerichs, R., and Greven, F. 2011. Estimating Nonresponse Bias and Mode Effects in a Mixed Mode Survey. International Journal of Market Research, 53: 669-686.

Maddala, G. S. (1986). Limited-Dependent and Qualitative Variables in Econometrics. Econometric Society Monographs, Cambridge University Press.

Pew Research Center. (2012). Assessing the Representativeness of Public Opinion Surveys. Report available at http://www.people-press.org/files/legacy-pdf/Assessing%20the%20Representativeness%20of%20Public%20Opinion%20Surveys.pdf

Peytchev, A. (2012). Multiple Imputation for Unit Nonresponse and Measurement Error. Public Opinion Quarterly, 76 (2): 214–237.

Presser, S., and Stinson, L. (1998). Data Collection Mode and Social Desirability Bias in Self-Reported Religious Attendance. American Sociological Review 63: 137-145.

Powers, J. R., Mishra, G., & Young, A. F. (2005). Differences in mail and telephone responses to self-rated health: use of multiple imputation in correcting for response bias. *Australian and New Zealand Journal of Public Health*, 29, 149–154.

Quigley, A. (2008). Adjustments for Mode Effect Bias for the Canadian Health Survey. Proceedings of the Survey Research Methods Section, American Statistical Association, Alexandria, VA.

Raglin, D., Zelenak, M.F., Davis, M.C., and Tancreto, J. 2008. Testing a New Field of Degree Question for the American Community Survey. DSSD American Community Survey Methods panel Memorandum Series, Chapter #ACS-MP-10.

Rao, J. N. K. and A. J. Scott (1981). The analysis of categorical data from complex sample surveys: Chi-squared tests for goodness of fit and independence in two-way tables. *The Journal of the American Statistical Association 76*, 221-230.

Rao, J. N. K., C. F. J. Wu, and K. Yue (1992). Some recent work on resampling methods for complex surveys. Survey Methodology 18 (2): 209-217.

Reiter, J. P. 2007. Small-sample degrees of freedom for multi-component significance tests with multiple imputation for missing data. Biometrika 94, 502–508.

Reiter, J. P., T. E. Raghunathan, and S. K. Kinney (2006). The importance of modeling the sampling design in multiple imputation for missing data. *Survey Methodology 32* (2), 143-149.

Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association 91* (434), 473-489.

Rubin, D. B. (2004). *Multiple Imputation for Nonresponse in Surveys* (Wiley Classics Library). Wiley-Interscience.

Soulakova, J., Davis, W.W., Hartmen, A., and Gibson, J. (2009). The Impact of Survey and Response Modes on Current Smoking Prevalence estimates Using TUS-CPS: 1992-2003. *Survey Research Methods*. 3: 123-137.

Tourangeau, R., and Smith, T. (1996). Asking Sensitive Questions: The Impact of Data Collection Mode, Question Format, and Question Context. Public Opinion Quarterly, 60: 275-304.

Train, K. and W. W. Wilson (2008). Estimation on stated-preference experiments constructed from revealed-preference choices. *Transportation Research Part B: Methodological 42* (3), 191-203.

Wooldridge, J. M. (2010). *Econometric Analysis of Cross Section and Panel Data.* 2nd ed. The MIT Press.

Voogt, R.J.J., and Saris, W.E. (2005). Missed Mode Designs: Finding the balance Between Nonresponse Bias and Mode Effects. *Journal of Official Statistics*, 21: 367-387.