How Hard Is it to Remove Mode Effects in Multimode Surveys? Basic Weighting v. Three Model-Based Methods

Matt Jans,[1] Randy ZuWallack,[1] Kelly Martin,[1] Thomas Brassell,[1] James Dayton,[1] Stephen Immerwahr,[2] Amber Levanon Seligson,[2] Sahnah Lim[3]
[1]ICF, Rockville, MD & Burlington, VT
[2]NYC Department of Health and Mental Hygiene (DOHMH), New York, NY
[3]New York University School of Medicine, New York, NY

Abstract
Multimode data collection is increasingly popular as survey funders and data collectors try to counteract declining response rates and increasing nonresponse bias risk. Using multiple modes can capture respondents who are different from those responding by the primary mode, thus reducing nonresponse bias. However, for long-running single-mode surveys, mode effects can introduce undesirable changes in trends. This presentation compares three mode adjustment methods (Kolenikov and Kennedy, 2014) with standard weighting using the 2017 New York City Social Determinants of Health survey. Three health outcomes that differed between random digit dial (RDD) and address-based samples (ABS) in unweighted analyses were the focus. Standard weighting removed mode differences for all three outcomes. Further adjustment with a regression method (RM), multiple imputation (MI) method, and an implied utility multiple imputation (IUMI) method moved estimates closer to weighted RDD estimates, which was considered the gold standard. These adjustments ranged from -1.41 to 1.35 percentage points beyond basic weighting. Results are discussed in the context of mode transition and implementation.

Key Words: multimode data collection; ABS; address-based sampling; mode adjustment; nonresponse; health statistics

1. Background and Objectives

Multimode data collection is increasingly popular as survey researchers address the challenges of continually-declining response rates and the related cost increases. Multimode data collection has several other benefits over single-mode surveys. Using multiple modes can capture respondents who are different from people who would respond in any single mode. That increase in overall response propensity can increase sample size relative to single-mode surveys. When respondents are offered a mode that matches their preferred mode, response propensity can increase dramatically (Olson, Smyth, & Wood, 2012).

However, multimode surveys are not without their challenges. One of the largest challenges impeding multimode survey adoption is the risk of mode effects: differences in responses or estimates due to the sampling or data collection mode. This concern is particularly relevant to long-running single-mode surveys because mode effects can lead to unexpected and undesirable changes in statistical trends over time, and make true population change difficult to ascertain.

This paper addresses this challenge by demonstrating several methods for adjusting multimode data to remove mode effects.

1.1 Mode Effects and Adjustment Methods

The literature on survey mode[1] and mode effects is vast (Aguinis, Sturman, & Pierce, 2007; Cernat, Couper, & Ofstedal, 2016; De Leeuw, 2005; Fowler, Roman, & Di, 1998; Johnson & Williams, 2013; Link, Battaglia, Frankel, Osborn, & Mokdad, 2008; Montaquila & Brick, 2012; Olson et al., 2012; Vannieuwenhuyze, Loosveldt, & Molenberghs, 2011; Vannieuwenhuyze, Loosveldt, & Molenberghs, 2012), but most survey designers and data users are primarily concerned with avoiding or removing mode effects from survey data and estimates. There are many ways that multiple modes can be incorporated into a survey design (see Kolenikov & Kennedy, 2014 for a review). This paper evaluates four methods: simple demographic weighting, regression-based adjustment, imputation-based adjustment, and a hybrid method that combines regression and imputation.

In most mode adjustment approaches, one mode is identified as the "adjustment standard" (i.e., the mode to which other modes will be adjusted). This decision is arbitrary but should be guided by the goals of the adjustment. For example, self-administered modes are often found to obtain more valid responses than interviewer-administered modes (e.g., Tourangeau & Yan, 2007). If the goal of the adjustment is reduce or remove the effect of interviewers on responses, then the self-administered mode would be the appropriate adjustment standard. However, long-running surveys aiming to avoid breaks in time series trends, such as a random digit dial (RDD) phone survey that is transitioning to address-based sampling with self-administration, will find it more useful to think of the historical mode as the adjustment standard. If changes in mode over time produce breaks in series, adjusting to the historical mode should avoid or reduce those breaks.

Broadly speaking, methods aimed at reducing mode effects are divided into methods based on logistic regression and methods based on imputation. Regression-based methods involve predicting responses to target survey questions from demographics and other predictors, including a term (i.e., predictor variable) for the mode of data collection in the regression model. This term essentially measures the mode effect, adjusted for other characteristics that contribute to producing the response, so that the predicted values are "mode-adjusted" and can be used to create estimates that are free of mode effects. Imputation-based methods approach the problem differently, treating the mode that is not the adjustment standard as missing. For example, in the RDD-to-ABS transition scenario, the data collected by ABS and self-administration would be treated as missing, and data from the "nonmissing" data (i.e., RDD phone interviews) would be used to predict missing ABS cases based on a set of predictors available on both samples.

[1] In this paper, "mode" refers to a combination of sampling frame and data collection mode (e.g., phone interviews collected from an RDD frame, or data collected by self-administration from an ABS frame).

1.2 Research Questions

This project had three research questions:

1. Can mode differences be removed with standard demographic weighting?
2. Do more intensive regression and imputation methods improve mode adjustments?
3. Does the extra effort for sophisticated adjustments result in higher quality data?

2. Methods

2.1 Social Determinants of Health Survey

The data used for this study come from the New York City Department of Health and Mental Hygiene's (NYC DOHMH) 2017 Social Determinants of Health (SDH) Survey, a collaboration between DOHMH and ICF. The survey used a dual-frame landline and cell phone RDD sample, and ABS. Computer-assisted telephone interviewing (CATI) was conducted with the RDD sample yielding 1,433 interviews with NYC adults, mostly in English and also in Spanish, Russian, Mandarin, and Cantonese. Sampled addresses from the ABS frame were randomly assigned to either mail-push-to-web with a mail survey nonresponse follow-up (3076 addresses), or to a mail survey with web nonresponse follow-up (3076 addresses). The ABS sample and self-administered modes yielded 902 completed questionnaires from NYC adults; about 2/3 of which were completed by mail. Data were collected between May and June, 2017.

2.2 Population Health Measures Assessed

For feasibility of this proof-of-concept, three population health measures that exhibited mode effects in prior research were used (Immerwahr, Lim, Brassell, et al. 2018). Table 1 describes the health estimates assessed and their source questions.

Table 1: Population Health Measures Assessed and their Definition[2]

Health Measure	*Definition*	*Question Wording*
Limited functioning	14 or more days in the past month	During the past 30 days, for about how many days did poor physical or mental health keep you from doing your usual activities, such as self-care, work, or recreation?
Poor mental health	14 or more days mental health was "not good" in the past month	Now thinking about your mental health, which includes stress, depression, and problems with emotions, for how many days during the past 30 days was your mental health not good?
Fair/poor health	Self-rated health is "fair" or "poor"	Would you say that in general your health is excellent, very good, good, fair or poor?

[2] Questions used come from health-related quality of life (HRQOL) questions used by the Centers for Disease Control and Prevention. More information can be found online at https://www.cdc.gov/hrqol/methods.htm

2.2 Weighting and Mode Adjustment Methods Evaluated

Using Kolenikov and Kennedy (2014) as a guide, weighting and mode adjustments proceeded in the following manner. First, each sample (RDD and ABS) was weighted independently of the other to NYC borough-level population control totals for age, gender, race/ethnicity, marital status, children present, and education. RDD interviews were also weighted to account for telephone sample-frame overlap and coverage. Second, the three mode adjustment methods described in Kolenikov and Kennedy (2014) were implemented: logistic regression, multiple imputation chained equations (MICE), and implied utility multiple imputation (IUMI).

Logistic Regression (REG) — Logistic regression was used to estimate the probability of each outcome variable as a function of mode and other covariates with the following model form:

$$\widehat{p_i} = 1/(1 + e^{-(\boldsymbol{\beta}' \boldsymbol{x_i} + \gamma m_i)})$$

The mode parameter is the estimated mode effect, which can be subtracted from the logit.

The mode-adjusted estimate is the sum of the probabilities, $\tilde{p} = \sum_i 1/(1 + e^{\boldsymbol{\beta}' \boldsymbol{x_i}})$.

Multiple Imputation Chained Equations (MICE) — This method treats the outcome for the secondary mode as missing and imputes it. First a logistic regression model estimates the probability of the outcome. Second, a random draw from a univariate distribution (u) determines whether to assign a "yes" or "no" (i.e., $y_i' = 1$ ($u \leq \hat{p}_i$) or $y_i' = 0$ ($u > \hat{p}_i$)) to the respondent.

The mode adjusted estimate is the average of the outcome including the imputations: $\tilde{p} = (\sum_i y_i' + \sum_i y_i)/n$.

This process was repeated for 50 imputations and then averaged over imputations using SAS PROC MI.

Implied utility multiple imputation (IUMI) — The IUMI is based on logistic regression, but for a utility model, in which the outcome is associated with an underlying latent variable:

$\boldsymbol{y_i} = \mathbf{1}$ (if $\boldsymbol{y_i^*} > \mathbf{0}$) or $\boldsymbol{y_i} = \mathbf{0}$ (if $\boldsymbol{y_i^*} \leq \mathbf{0}$),

where,

$$\boldsymbol{y_i^*} = \boldsymbol{\beta}' \boldsymbol{x_i} + \boldsymbol{\gamma m_i} + \boldsymbol{\varepsilon_i}$$

is the underlying latent utility associated with $\boldsymbol{y_i}$.

Kolenikov and Kennedy developed a mode adjustment by simulating values of the underlying utility, $\boldsymbol{y_i^*}$, and removing the mode effect: $\boldsymbol{y_i^{(m)}} = \boldsymbol{y_i^*} - \boldsymbol{\gamma m_i}$. The method simulates $\hat{\boldsymbol{\varepsilon}}_{\boldsymbol{i}}$ from a truncated logistic distribution conditional on the observed response $\boldsymbol{y_i}$. That is, the error term $\boldsymbol{\varepsilon_i}$ is constrained based on the observed response

$\boldsymbol{y_i}$, $\boldsymbol{\varepsilon_i} > -(\boldsymbol{\beta' x_i} + \boldsymbol{\gamma m_i})$ when $\boldsymbol{y_i} = \mathbf{1}$ $(\boldsymbol{y_i^*} > \mathbf{0})$ and $\boldsymbol{\varepsilon_i} \leq -(\boldsymbol{\beta' x_i} + \boldsymbol{\gamma m_i})$ when $\boldsymbol{y_i} = \mathbf{0}$ $(\boldsymbol{y_i^*} \geq \mathbf{0})$. The mode adjusted estimate is the average of the simulations: $\tilde{\boldsymbol{p}} = \sum_i \boldsymbol{y_i^{(m)}}$.

This process was repeated for 50 simulations and then averaged over all simulations.

For the purposes of mode adjustment, we considered the data collected by RDD to be the adjustment standard because DOHMH has the most experience collecting its annual adult health survey using that mode.

Each method was estimated on unweighted data, with the weighting dimensions described above included as predictors to obtain similar results as if we had implemented the techniques using survey weights. Due to the exploratory nature of this research, we used a combination of statistical testing and simple magnitude review to assess the impact of adjustments.

3. Results

3.1 RDD v. ABS Differences

Figure 1 shows the differences between RDD and ABS for unweighted and weighted survey data. Difference is calculated as RDD – ABS; a positive value indicates higher incidence in the RDD sample. The differences between RDD and ABS were not significant after weighting, and the reduction in magnitude of those differences suggests that weighting alone removed most of the mode effect for limited functioning and fair/poor health. It did not remove the mode effect for poor mental health.

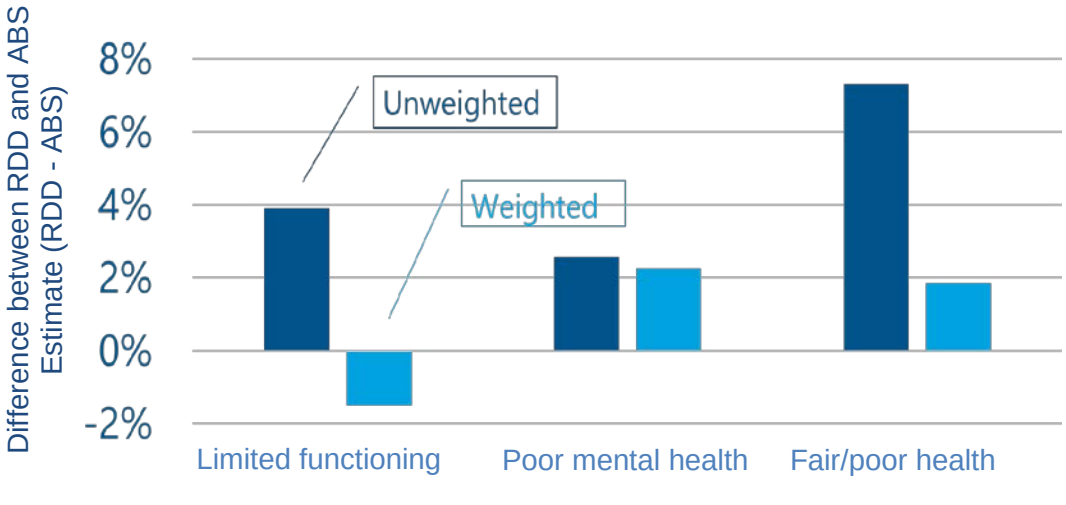

Figure 1: Mode differences removed with standard demographic weighting

3.2 RDD-only v. Mode-adjusted

Figure 2 shows the differences between estimates from the RDD sample and mode-adjusted estimates by each adjustment method. The difference was calculated as the mode-adjusted estimate minus the RDD-only estimate so that a positive value indicates higher incidence after mode-adjustment. The first thing to note is the small scale on which all adjustments fall, ranging from 0.5% to -0.8%, showing again that even weighting removes much of the mode effect. The second observation, based on review of the magnitudes of differences between modes in Figure 2, is that the regression, imputation (MICE), and IUMI remove even more of the difference between modes, with regression appearing to remove more of that difference than either MICE or IUMI across health indicators. Mode effects on poor mental health and fair/poor health appear to be more difficult to remove than mode effects on limited functioning.

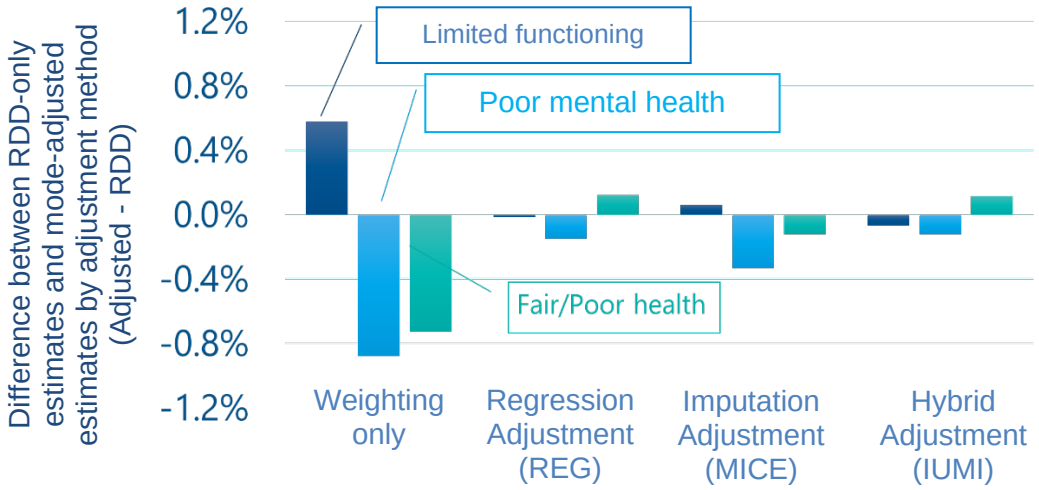

Figure 2: Change in estimate from RDD-only to mode-adjusted by adjustment method

3.3 Single-source v. Combined Estimates

Table 2 shows combined and single-source (e.g., RDD or ABS) estimates to show expected changes in trends that would occur from using an ABS-only approach versus a combined approach. After demographic weighting, there were no significant differences between the ABS and RDD. However, lack of statistical differences may be a result of high standard errors. Thus, mode effects may still exist. For example, the 2.3 percentage point difference between RDD and ABS in the measure of poor mental health may have become significant if the sample size had been larger. It is noteworthy that ABS-only point estimates were higher than either RDD or combined estimates for limited functioning and poor mental health but were lower for fair or poor general health. In general, the combined estimates were closer to RDD-only than ABS-only estimates.

Table 2: Health Estimates and 95% Confidence Intervals for Each Mode and Combined

Health Estimate	*Combined*	*RDD*	*ABS*	*Difference (RDD-ABS)*
Limited functioning	10.9% (±1.9%)	10.3% (±2.0%)	11.8% (±3.5%)	-1.5% (±4.1%)
Poor mental health	6.5% (±1.4%)	7.4% (±1.9%)	10.9% (±1.8%)	2.3% (±2.7%)
Fair/Poor health	19.1% (±2.3%)	19.8% (±2.8%)	17.9% (±4.0%)	1.9% (±4.9%)

3.4 Effect of Adjustment on Root Mean Square Error (RMSE)

Table 3 displays the impact of each method on root mean squared error (RMSE) as an indicator of the overall impact on bias and variance. The RMSE is based on the variance of the mode-adjusted estimate and the squared difference between the mode-adjusted estimate and the RDD estimate. There is little variability between the methods in terms of RMSE.

Table 3: Effect of Adjustment on Root Mean Squared Error (SE^2 + $difference^2$)

Health Estimate	*Weights*	*Regression*	*Multiple Imputation*	*Implied Utility Multiple Imputation*
Limited functioning	1.1%	1.0%	1.2%	1.2%
Poor mental health	1.1%	0.9%	1.0%	1.0%
Fair/Poor health	1.4%	1.3%	1.6%	1.6%

4. Conclusions and Discussion

This paper presented four ways to adjust for mode effects in population surveys, particularly those that are long-running single-mode surveys. The results provide initial evidence to answer the following research questions.

1. *Can mode differences be removed with standard demographic weighting?*

   For two of the three health indicators assessed (limited functioning and fair/poor health), weighting alone removed most of the differences between RDD and ABS estimates. This was not the case for poor mental health.

2. *Do more intensive regression and imputation methods improve mode adjustments?*

   Each of the methods tested reduced the mode effect further. Regression adjustment and IUMI seemed to reduce it the most, but there was variability across estimates.

3. *Does the extra effort for sophisticated adjustments result in higher quality data?*

   In settings where the researcher has the time, resources, and motivation to develop estimate-specific, optimized mode adjustments, the answer is probably yes. However, some adjustments will come at the cost of root mean square error. Based on this research, we think that pursuing regression-based adjustment is worthwhile in many situations. It reduced the mode effects more than weights alone, and led to lower root mean square error than any other approach, including weighting. However, when there is limited time and resources, weighting may be an adequate approach to reducing bias due to mode effect.

The largest limitations of this study are the limited number of mode effect adjustments that were assessed. There are other mode effect adjustments that could be evaluated. Further, the current methods tested could be implemented on other health outcomes to assess their generalizability across estimates. This would help identify whether specific estimates or types of estimates (e.g., physical health, mental health, health care access and use) respond better to certain mode adjustment methods than others. Our next steps will be focused on evaluating these methods on other SDH health outcomes, and we encourage readers to test them on other surveys as well.

References

Aguinis, H., Sturman, M. C., & Pierce, C. A. (2007). Comparison of Three Meta-Analytic Procedures for Estimating Moderating Effects of Categorical Variables. *Organizational Research Methods*, *11*(1), 9–34. https://doi.org/10.1177/1094428106292896

Cernat, A., Couper, M. P., & Ofstedal, M. B. (2016). Estimation of Mode Effects in the Health and Retirement Study Using Measurement Models. *Journal of Survey Statistics and Methodology*, *4*(4), 501–524. https://doi.org/10.1093/jssam/smw021

De Leeuw, E. D. (2005). To mix or not to mix data collection modes in surveys. *Journal of Official Statistics*, *21*(5), 233–255.

Fowler, F. J., Roman, A. M., & Di, Z. X. (1998). Mode effects in a survey of Medicare prostate surgery patients. *Public Opinion Quarterly*, *62*(1), 29.

Immerwahr, S., Lim, S., Brassell, T., ZuWallack, R., Levanon Seligson, A. *Is ABS More Representative than RDD for Public Health Surveillance Surveys?* American Association for Public Opinion Research Conference, May 2018. Denver, CO. Unpublished conference paper, 2018

Johnson, P., & Williams, D. (2013). Comparing ABS vs. Landline RDD Sampling Frames on the Phone Mode. *Survey Practice*, *3*(3). Retrieved from http://www.surveypractice.org/index.php/SurveyPractice/article/view/251

Kolenikov, S., & Kennedy, C. (2014). Evaluating Three Approaches to Statistically Adjust for Mode Effects. *Journal of Survey Statistics and Methodology*, *2*(2), 126–158. https://doi.org/10.1093/jssam/smu004

Link, M. W., Battaglia, M. P., Frankel, M. R., Osborn, L., & Mokdad, A. H. (2008). A Comparison of Address-Based Sampling (ABS) Versus Random-Digit Dialing (RDD) for General Population Surveys. *Public Opinion Quarterly*, *72*(1), 6–27. https://doi.org/10.1093/poq/nfn003

Montaquila, J. M., & Brick, J. M. (2012, August). *Transitioning from RDD to ABS with mail as the primary mode*. Presented at the Joint Statistical Meetings, San Diego, California.

Olson, K., Smyth, J. D., & Wood, H. M. (2012). Does Giving People Their Preferred Survey Mode Actually Increase Survey Participation Rates? An Experimental Examination. *Public Opinion Quarterly*, *76*(4), 611–635. https://doi.org/10.1093/poq/nfs024

Tourangeau, R., & Yan, T. (2007). Sensitive questions in surveys. *Psychological Bulletin*, *133*(5), 859–883.

Vannieuwenhuyze, J., Loosveldt, G., & Molenberghs, G. (2011). A Method for Evaluating Mode Effects in Mixed-mode Surveys. *Public Opinion Quarterly*, *74*(5), 1027–1045. https://doi.org/10.1093/poq/nfq059

Vannieuwenhuyze, J. T., Loosveldt, G., & Molenberghs, G. (2012). A Method to Evaluate Mode Effects on the Mean and Variance of a Continuous Variable in Mixed-Mode Surveys. *International Statistical Review*, *80*(2), 306–322.