INVESTIGATING THE BIAS OF ALTERNATIVE STATISTICAL INFERENCE METHODS IN MIXED-MODE SURVEYS

Dissertation Research

06/02/2013

Z. Tuba Suzer Gurtekin

Advisors:

Steven G. Heeringa

Richard Valliant

PROGRAM IN SURVEY METHODOLOGY
INSTITUTE FOR SOCIAL RESEARCH

UNIVERSITY OF MICHIGAN

426 THOMPSON STREET
ANN ARBOR, MI 48106-1248
734-647-0038 FAX: 734-764-8263
michpsm@isr.umich.edu

Outline

- Motivations behind Mixed-Mode Surveys
- Typical Assumption in Mixed Mode Surveys: All modes produce correct data (no Mode Effects)
  - Mode effects confounded with mode choice
  - Existing Methods
- Proposed Imputation Methods to assess and adjust for mode effects
- Simulation Study Results
- Conclusions and Current Research

Mixed-Mode Surveys - Motivations

- Decreasing response rates (Curtin, Presser, & Singer, 2005; de Leeuw, 2005; de Leeuw & de Heer, 2002; Steeh, Kirgis, Cannon, & DeWitt, 2001)
- Increasing survey costs (Groves & Heeringa, 2006)
- Better understanding of measurement properties (Tourangeau & Smith, 1996)
- Trends in technology use
  - 17% of cell-owner adults use their cell-phones to go online in the U.S. (Pew Center, Cell Internet Use Survey, 2012)
  - Increasing trends in computer use, Internet access and Broadband Internet access rates (U.S. Census Bureau, 2011)

Mode Effects

- **Single-mode surveys:**
  - Differences in overall results: Mode effects are part of trade-off analysis, no assumption about the ignorability of mode effects
- **Mixed-mode surveys:**
  - Assumption: No mode effects
    - Social desirability bias: Respondents are more likely to misreport their statuses on sensitive topics conditioned on their status in the presence of an interviewer (Tourangeau & Smith, 1996; Tourangeau & Yan, 2007)
    - In-person respondents may be more immune to social desirability tendencies (Holbrook, Green, & Krosnick, 2003)
    - E.g., Income is a sensitive topic in the U.S. (Moore, Stinson & Welniak, Jr, 2000), also an observable characteristic

Mode Effects in Mixed-Mode Surveys

- Mode choice: Nonrandomized Mode Assignment
  - E.g., Respondents with higher education are more likely to respond in telephone mode than in in-person compared to respondents who have less than a 12th Grade education (CPS, March 2012)

- Mode effects are confounded by mode choice in mixed-mode surveys

Existing Methods to Assess Mode Effects

- Randomization and control other error sources (Jäckle, Roberts, & Lynn, 2010; Biemer, 2001)
  - Assign modes randomly
- Comparison to a single mode survey (Vannieuwenhuyze, Loosveldt, & Molengberghs, 2010; 2012)
  - Mixture distribution
  - Representativity assumption
  - Limited to two modes

Existing Methods to Adjust for Mode Effects

- Calibrate the mode proportions to fixed proportions (Buelens & VandenBrakel, 2011)
  - Include mode in the calibration estimator
  - Does not eliminate bias, instead aim to calibrate bias to yield unbiased change estimates
- Selection models (Cobben, 2009; Cobben, Schouten, & Bethlehem, 2006)
  - Include the sequential nature of mode choice in nonresponse weights

Measurement Error Model for a Mixed-Mode Survey

$$y_j = \mu(X_j, \beta^{(\mu)}) + R_{jT} B_{jT} + R_{jI} B_{jI} + \varepsilon_j$$

Mode choice

R_{jT}, R_{jI} : indicator variables for response mode

$\mathrm{E}_R(R_{jT}) = g(X_j; \psi) \equiv g_j$

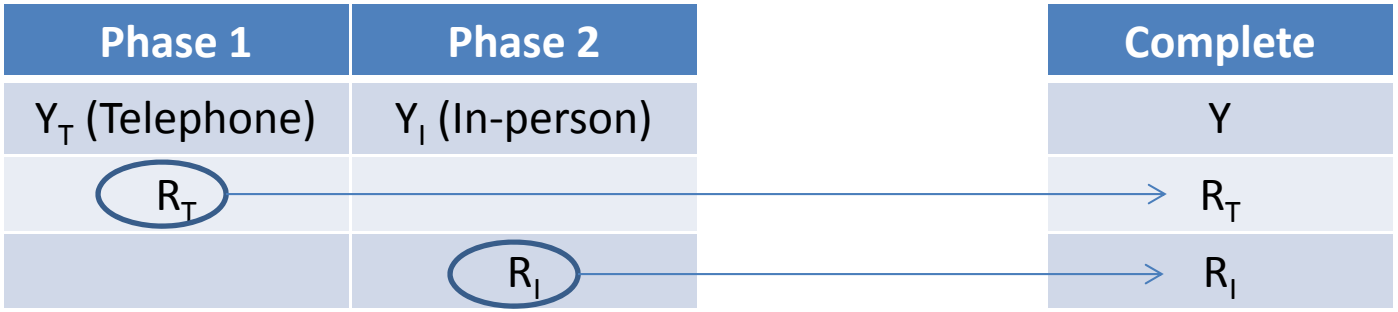
Mode effects

$\mu_j = \mu(X_j, \beta^{(\mu)})$

B_{jT}, B_{jI} : mode effects

Alternatively, $B_{jT} = Z_j^T B_T$, $B_{jI} = Z_j^T B_I$

Population Mean Ignoring Mode Effects



$$\bar{Y}_0 = \frac{1}{N}\left[\sum_{j\in U_T} y_j + \sum_{j\in U_I} y_j\right]$$

U_T = Telephone respondents

U_I = Inperson respondents

$$\mathrm{E}_M[\bar{Y}_0 - \bar{Y}] = \overline{(gB)}_U$$

a weighted average of telephone and in-person mode effects

Alternatively, Multiple Imputation Method

Phase 1	Phase 2
Y_T (Telephone)	Y_I (In-person)
R_T	NR_I
NR_T	R_I

$$\bar{Y}_T^*$$

$$\bar{Y}_T^* = \frac{1}{N}[\sum_{j \in U_T} y_j + \sum_{j \in U_I} y_{jT}^*]$$

$$\mathrm{E}_M \mathrm{E}_I [\bar{Y}_T^* - \bar{Y}] = B_{jT}$$

$$\bar{Y}_I^*$$

$$\bar{Y}_I^* = \frac{1}{N}[\sum_{j \in U_I} y_j + \sum_{j \in U_T} y_{jI}^*]$$

$$\mathrm{E}_M \mathrm{E}_I [\bar{Y}_I^* - \bar{Y}] = B_{jI}$$

How to Combine Mode-Specific Estimates?

$$\bar{Y}^{*} = \alpha \bar{Y}_{T}^{*} + (1-\alpha)\bar{Y}_{I}^{*} \qquad 0 \le \alpha \le 1$$

Empirical Alternative Combination Methods:

Method 1 (CM_1) – Simple average estimator:

$$\alpha = \frac{1}{2}$$

Method 2 (CM_2) – Weighted inversely according to the variances of the estimated means :

$$\alpha = \frac{1}{Var(\bar{Y}_{P}^{*})} \Big/ \sum_{P} \frac{1}{Var(\bar{Y}_{P}^{*})}$$

Method 3 (CM_3) – Weighted inversely according to the mean square errors of the estimated means:

$$\alpha = \frac{1}{MSE(\bar{Y}_{P}^{*})} \Big/ \sum_{P} \frac{1}{MSE(\bar{Y}_{P}^{*})}$$

Evaluate Alternative Combination Methods and Competing Method

$$\text{RelBias}_{CM_l} = \left(\frac{\bar{Y}^* - \bar{Y}}{\bar{Y}^*} \right)$$

where *l=1,2,3* is the combination methods

$$\text{RelBias}_{MODES\_IGNORED} = \left(\frac{\bar{Y}_0 - \bar{Y}}{\bar{Y}_0} \right)$$

Simulation Study Description

- Simulation study: Total Family Income
  - Create hypothetical populations using Current Population Survey (CPS), 1973, and Social Security Records: Exact Match Data
- CPS March Supplement
  - Rotating panel survey
  - Produces data on the U.S. labor force
  - The rotation scheme follows a 4-8-4 pattern
  - The majority of first and fifth waves are in-person interviews
  - For the other waves, respondents are given the choice to do the interview by telephone or in-person visits
  - Majority of interviews from the other waves are telephone

Simulation Study Description

- Total Family Income is constructed by summing up eight income types as reported in CPS over the household head and spouse
- The data exclude the records with item missing in any of the CPS income type and CPS Total Family Income
- Top-coded income values are excluded
- Adjusted Gross Income (AGI) as matched from IRS records used as benchmarks
  - Since Received Welfare Amount is not reported in as part of AGI, a control variable is used in the models to reflect the differences in the income constructs

Simulation Study Description

- X covariates: Race-ethnicity, Living Quarters Type, Region, Industry Type, Job Type, Spouse Work Status, Presence of Children, Respondent Status of Householder

- Regression analysis suggests that there are not differences between modes for this subset
  - Distribution is skewed for whites (94%), laborer (5%)

Hypothetical Populations

Varying Mode Effects:

$$Y_{Ij} = \beta^{(AGI)} Y_j^{(AGI)}$$

$$Y_{Tj} = \beta^{(AGI)} Y_j^{(AGI)}$$

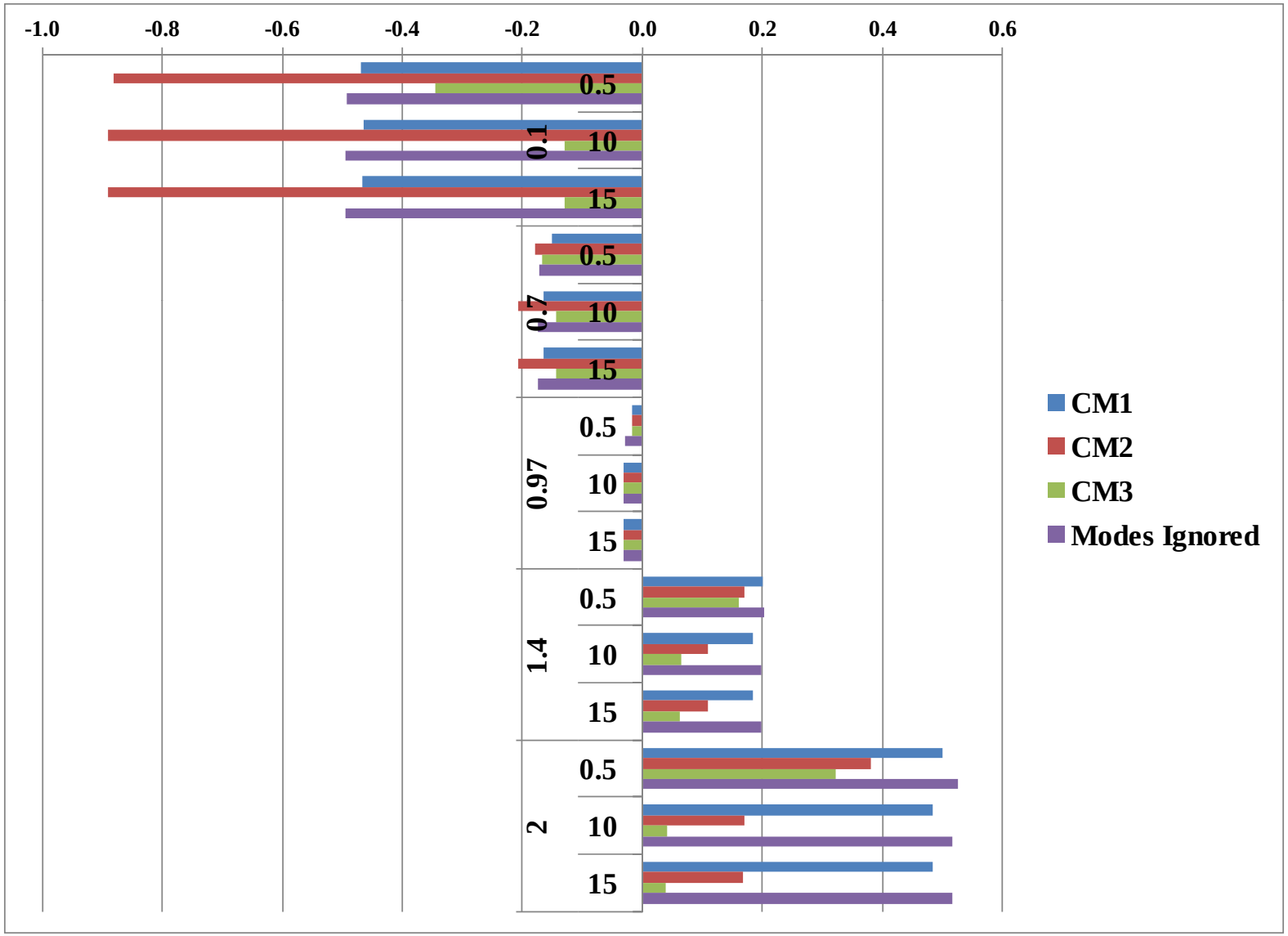
- Beta constant for in-person varies between 0.1-2.0 based on AGI
- Corresponds to Relative Bias of (-0.9 to 1)

Varying Goodness of Model Fit :

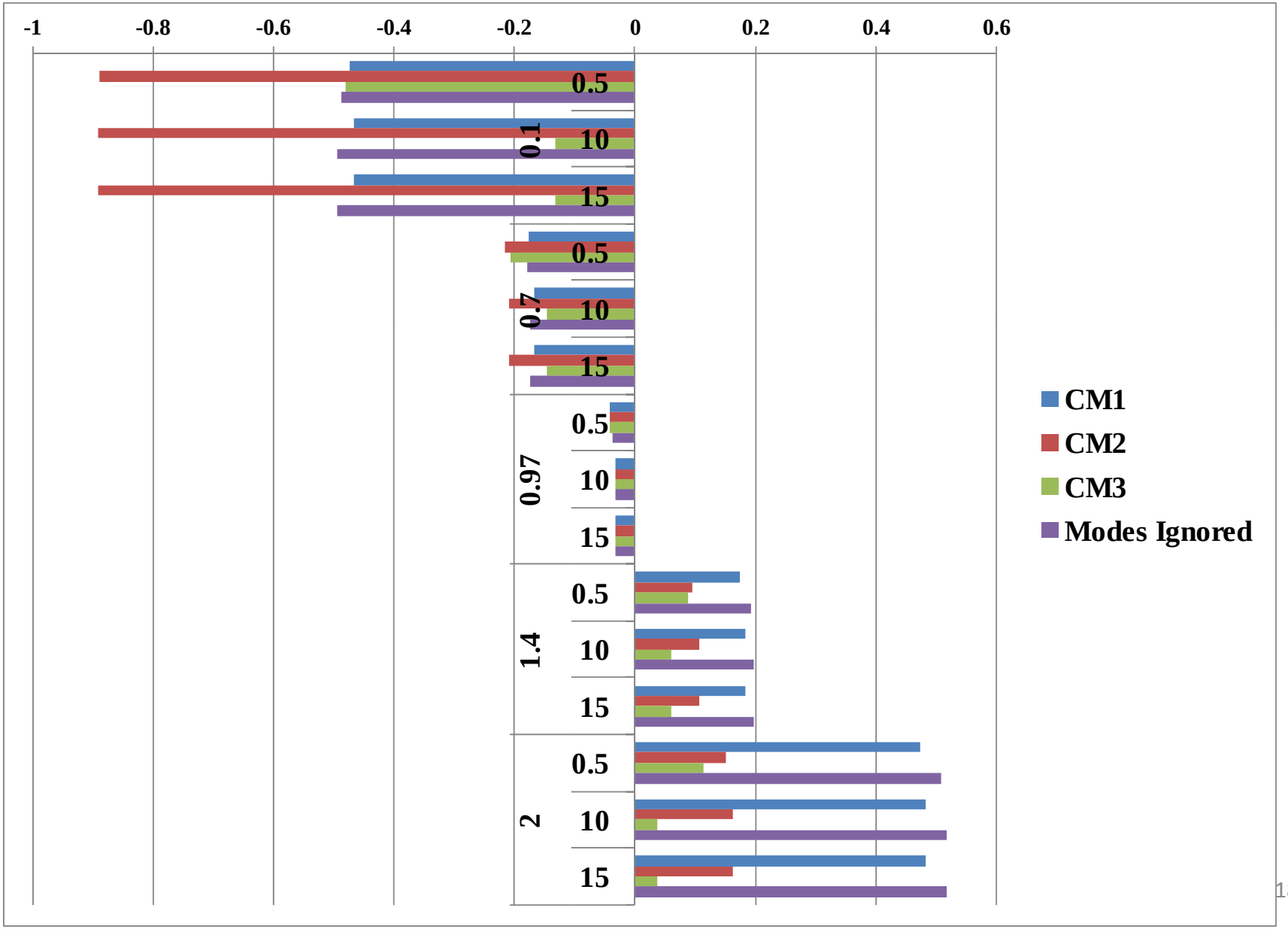
$$\hat{Y}_j = X_j^{(Y)} \hat{\beta}^{(Y)} + e_j$$

$$e_j \overset{iid}{\sim} N\left(0, \sigma^2 / c\right) \quad , \quad c \in (0.5, 10, 15)$$

Simulation Study Results: Relative Biases, Fixed Mode Choice



Simulation Study Results: Relative Biases, Variable Mode Choice



Conclusions and Current Research

- Possible severe bias in traditional method
- Evaluation of model assumptions
  - Feasibility
- Alternatively, sensitivity analyses can be conducted in the absence of benchmarks

Thank you.

Contact information: tsuzer@umich.edu

References

Biemer, P. (2001). Nonresponse bias and measurement bias in a comparison of face to face and telephone interviewing. Journal of Official Statistics, 17(2), 295-320.

Buelens, B., & Van den Brakel, J. (2011). Inference in Surveys with Sequential Mixed-Mode Data Collection. Statistics Netherlands.

Cobben, F. (2009). Nonresponse in Sample Surveys Methods for Analysis and Adjustment. cbs.nl. Universiteit van Amsterdam.

Cobben, F., Schouten, B., & Bethlehem, J. (2006). A Model for Statistical Inference based on Mixed Mode Interviewing. European Conference on Quality in Survey Statistics.

Curtin, R., Presser, S., & Singer, E. (2005). Changes in telephone survey nonresponse over the past quarter century. Public Opinion Quarterly, 69(1), 87-98.

De Leeuw, E. D. (2005). To mix or not to mix data collection modes in surveys. Journal of Official Statistics, 21(2), 233-255.

De Leeuw, E., & De Heer, W. (2002). Trends in household survey nonresponse: A longitudinal and international comparison. In R. M. Groves, D. A. Dillman, J. L. Eltinge, & R. J. A. Little (Eds.), Survey nonresponse (pp. 41-54). New York: Wiley series in probability and statistics.

References

Dillman, D., & Tarnai, J. (1988). Administrative issues in mixed mode surveys. In R. M. Groves, P. P. Biemer, L. E. Lyberg, J. T. Massey, W. L. Nicholls II, & J. Waksberg (Eds.), Telephone Survey Methodology (pp. 509-528). Wiley Series in Survey Methodology.

Greenlees, J. S., Reece, W. S., & Zieschang, K. D. (1982). Imputation of missing values when the probability of response depends on the variable being imputed. Journal of the American Statistical Association, 77(378), 251-261.

Groves, R. M., & Couper, M. P. (1998). Nonresponse in household interview surveys. Wiley Series in Probability and Statistics Survey Methodology Section.

Groves, R. M., & Heeringa, S. G. (2006). Responsive design for household surveys: tools for actively controlling survey errors and costs. Journal of the Royal Statistical Society. Series A. Statistics in society, 169(3), 439-457.

Holbrook, A. L., Green, M. C., & Krosnick, J. A. (2003). Telephone versus face-to-face interviewing of national probability samples with long questionnaires: Comparisons of respondent satisficing and social desirability response bias. *Public Opinion Quarterly*, *67*(1), 79.

Jäckle, A., Roberts, C., & Lynn, P. (2010). Assessing the effect of data collection mode on measurement. International Statistical Review, 78(1), 3–20.

Knauper, B., & Schwarz, N. (2004). Why Your Research May Be Out of Order How Age-sensitive Context Effects May Lead As Astray. Psychologist, 17(1), 28-31.

References

Körmendi, E. (1988). The quality of income information in telephone and face to face surveys. In R. M. Groves, P. P. Biemer, L. E. Lyberg, J. T. Massey, W. L. Nicholls II, & J. Waksberg (Eds.), Telephone Survey Methodology (pp. 341-356). Wiley Series in Survey Methodology.

Krosnick, J., & Alwin, D. (1987). An evaluation of a cognitive theory of response-order effects in survey measurement. Public Opinion Quarterly, 51, 201-219.

Little, R.J.A., & Rubin, D.B. (2002). Statistical analysis with missing data (2nd ed.). Hoboken, New Jersey: Wiley Series in Probability and Statistics.

Rubin, D. B. (1987). Multiple imputation for nonresponse in surveys. Hoboken, New Jersey: Wiley Classics Library.

Steeh, C., Kirgis, N., Cannon, B., & DeWitt, J. (2001). Are They Really as Bad as They Seem? Nonresponse Rates at the End of the 20th Century. Journal of Official Statistics, 17(2), 227-247.

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

U.S. Census Bureau (2011). Exploring the Digital Nation. Computer and Internet Use at Home.

References

Vannieuwenhuyze, J., Loosveldt, G., & Molenberghs, G. (2010). A Method for Evaluating Mode Effects in Mixed-mode Surveys. Public Opinion Quarterly, 74(5), 1027-1045.

Vannieuwenhuyze, J., 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.

Bias of Sample Mean in Mixed-Mode Surveys

Measurement error model:

$$y_i = \mu_i + B_{p,g} + \varepsilon_i,\ \left(i \in U_{p,g},\ \varepsilon_i \sim (0, \sigma_\varepsilon^2)\right)$$

Sample mean:

$$\bar{y} = \sum_{i=1}^{n} y_i \Big/ n$$

The bias of $\bar{y}$:

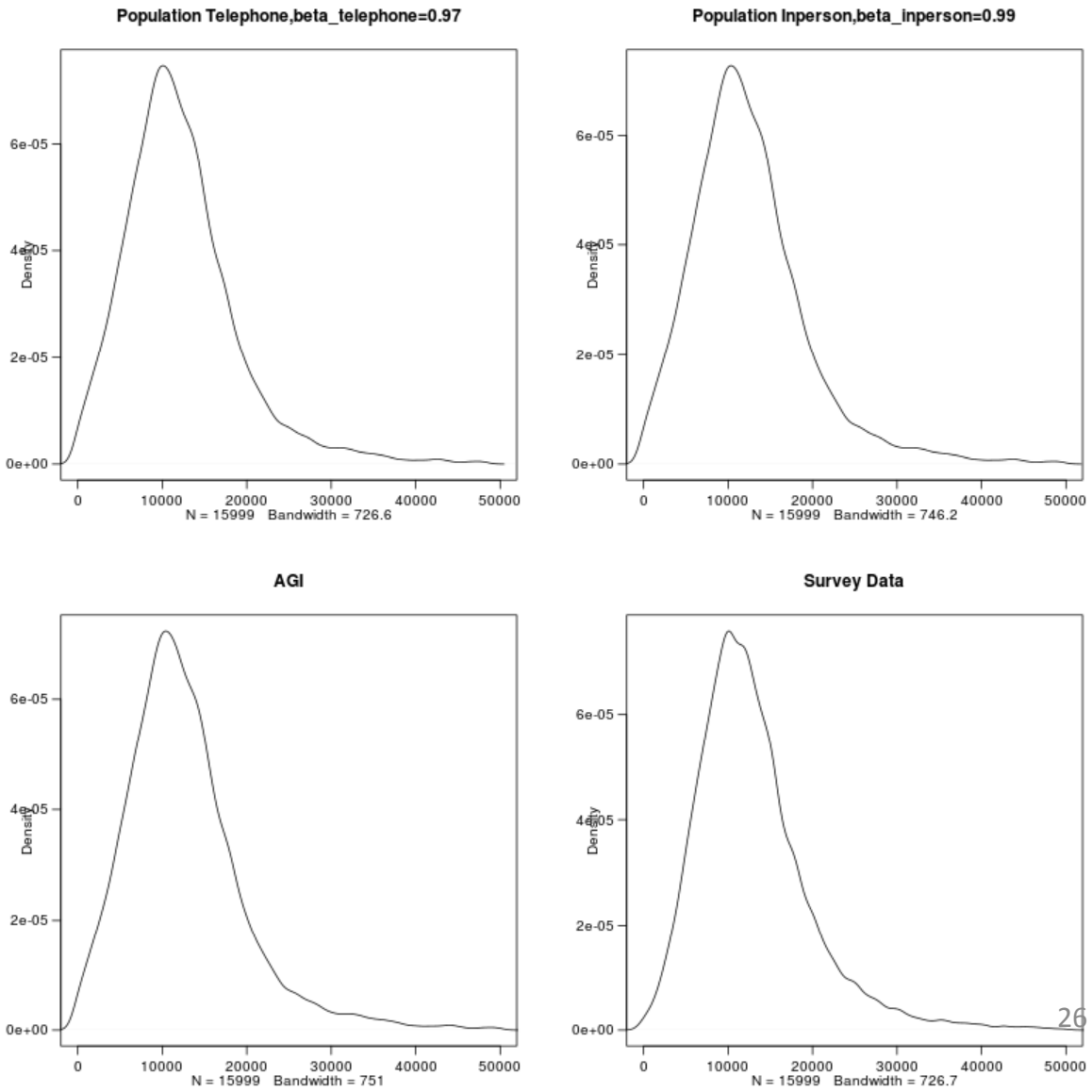
$$\sum_g P_{p,g} B_{p,g}$$

Simulation Study: Total Family Income Hypothetical Populations - Varying mode effects

$$Y_{iP} = \beta AGI_i$$

$$\begin{bmatrix} \beta_{Telephone} \\ \beta_{Inperson} \end{bmatrix} = \begin{bmatrix} 0.97 \\ (0.1, 0.97, 2.0 \end{bmatrix}$$

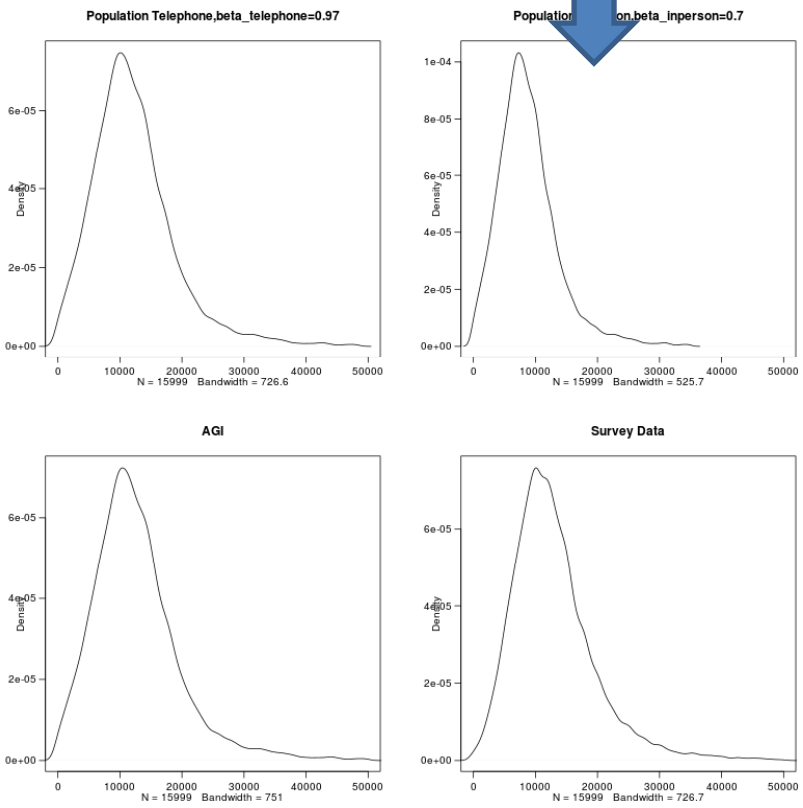
where *i* represents a respondent



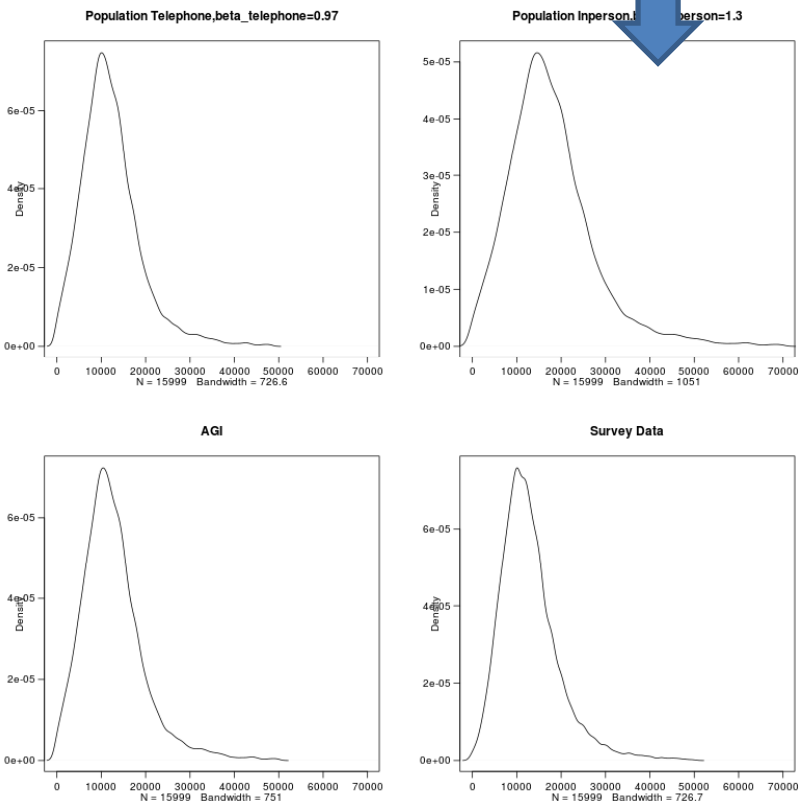
Simulation Study: Total Family Income

Hypothetical Populations - Varying mode effects

beta_inperson=0.7



beta_inperson=1.4



Simulation Results – Evaluation of Combination Methods

- Including item missing in imputation yields
  - Larger absolute relative bias on the average
  - Larger variation

- Combination method 3 outperforms the competing method using deterministic regression model, but not in stochastic regression model simulations

Imputation Model: Ignorable Mode Effects - Continuous Variables

Normal Linear Regression Model:

$$Y_{jT} \sim N(X_j \beta, \sigma^2)$$

Assuming the standard noninformative prior distribution

$$\Pr(\beta, \sigma \mid X) \propto \frac{1}{\sigma^2} \rightarrow \qquad (\beta \mid \sigma^2, y) \sim MVN(\hat{\beta}, V(\hat{\beta})\sigma^2)$$

where, $\hat{\beta} = (X^T X)^{-1} X^T Y$,and $V(\hat{\beta}) = (X^T X)^{-1}$

$$\Pr(\sigma^2 \mid y) \sim Inv - \chi^2(n-k, s^2)$$,and

$$s^2 = \frac{1}{n-k}(y - X\hat{\beta})^T (y - X\hat{\beta})$$

where n is sample size and k number of parameters

Imputation Model: Nonignorable - Selection Models – Continuous Variables

A model for the mode choice mechanism:

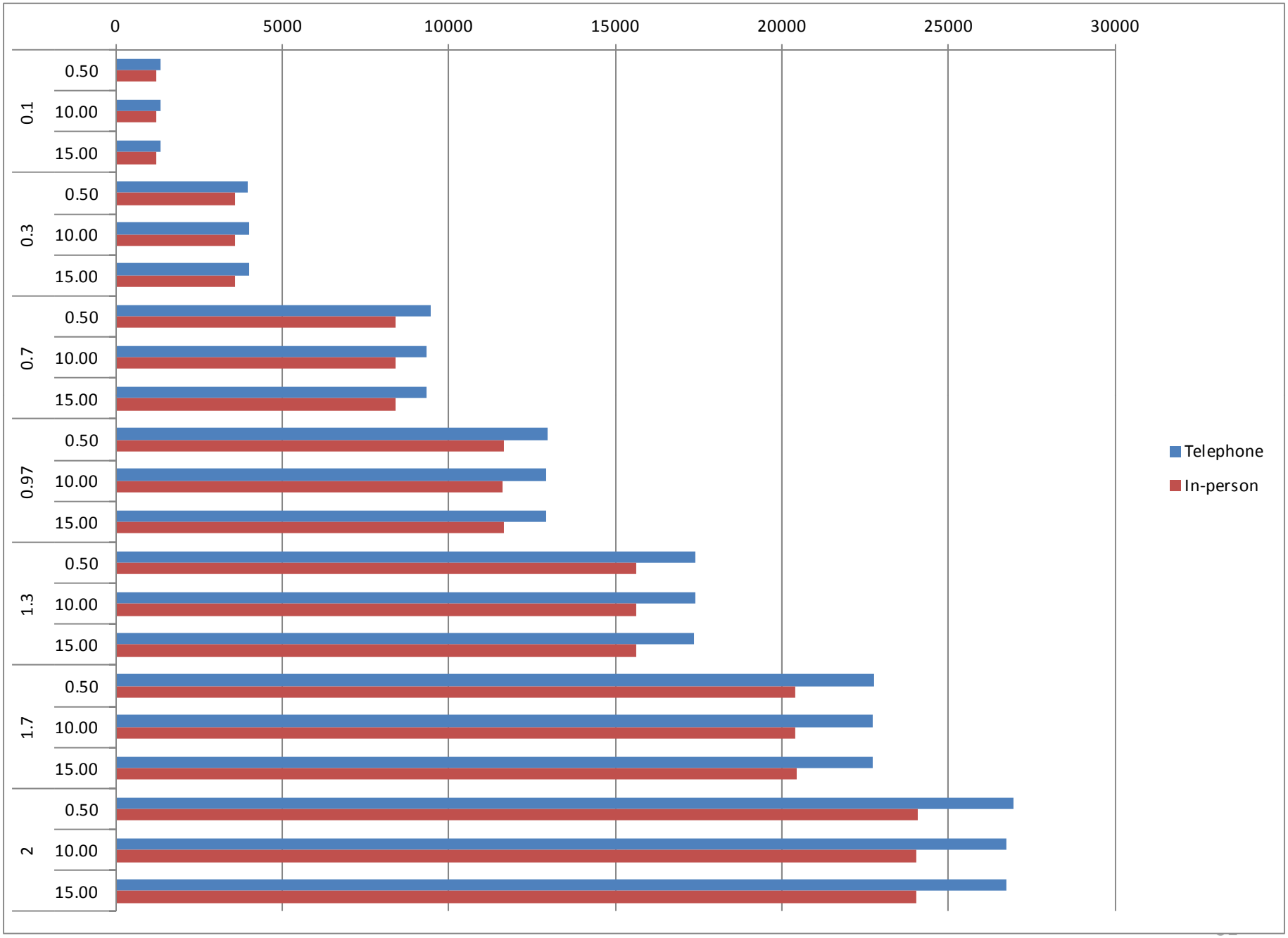
$$\Pr(R_{jT} = 0 \mid X_j^R, Y_{jT}; \psi) = \left[1 + \exp(-X_j^{(R)}\beta^{(R)} - \gamma Y_j)\right]^{-1}$$

A complete data model:

$$(Y_{jT} \mid X_j^{(Y)}; \theta) \sim N(X_j^{(Y)}\beta^{(Y)}, \sigma^2)$$

$$L_{full}(\theta, \psi \mid Y_T, RjT) = \prod_{j \in U_T} \frac{1}{\left[1 + \exp(-X_j^{(R)}\beta^{(R)} - \gamma Y_j)\right]} \frac{1}{\sigma} \Phi\left(\frac{Y_j - X_j^{(Y)}\beta^{(Y)}}{\sigma}\right) \times$$

$$\prod_{j \in U_R} \int_{-\infty}^{\infty} \left(1 - \frac{1}{\left[1 + \exp(-X_j^{(R)}\beta^{(R)} - \gamma Y_j)\right]}\right) \frac{1}{\sigma} \Phi\left(\frac{Y_j - X_j^{(Y)}\beta^{(Y)}}{\sigma}\right) dy_j$$


0
5000
10000
15000
20000
25000
30000
0.1
0.3
0.7
0.97
1.3
1.7
2
0.50
10.00
15.00
Telephone
In-person

Subset of Current Population Survey (CPS), 1973, and Social Security Records: Exact Match Data

The mode distribution shifts across the month in sample as expected (n=15,999)

Mode	M1	M2	M3	M4	M5	M6	M7	M8
Telephone	2%	34%	60%	66%	6%	53%	63%	65%
In-person	98%	66%	40%	34%	94%	47%	37%	35%

Overall: 56% in-person, 44% telephone

Hypothetical Populations Income Means- Varying mode effects

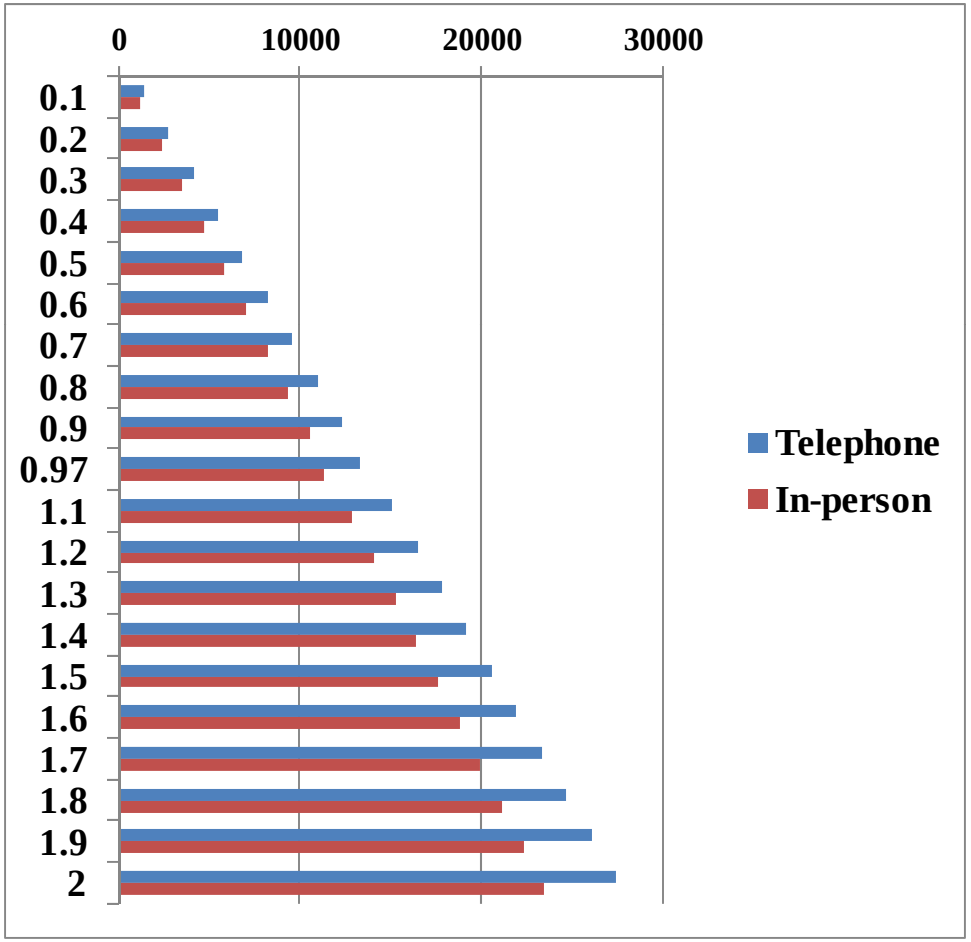
$$Y_{Ij} = \beta^{(AGI)} Y_j^{(AGI)}$$

$$Y_{Tj} = \beta^{(AGI)} Y_j^{(AGI)}$$

Beta constant for in-person varies between 0.1-2.0 based on AGI

Corresponds to Relative Bias of (-0.9 to 1)

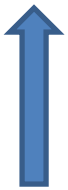
The higher mean for telephone respondents is preserved



Simulation study 1: Household Income

Varying goodness of model fit

$$\hat{Y}_j = X_j^{(Y)} \hat{\beta}^{(Y)} + e_j \qquad e_j \overset{iid}{\sim} N\left(0, \sigma^2 / c\right)$$



c(0.5,10,15)

where *i* represents a respondent

Alternatively, Multiple Imputation Method

Phase 1	Phase 2
Y_1 (Telephone)	Y_2 (In-person)
R_T	NR_I
NR_T	R_I

- A special case of a missing data problem
- Impute data for each phase through a series of multiple imputation models as if all units had reported in that particular mode
- Impute nonrespondent data for Telephone and In-person phases via multiple imputation models
- *X* covariates in the models are combination of personal and residential data (such as age, gender, etc.)
- Continuous variable: Normal linear regression model, noninformative prior distribution.