Institute of Education Sciences
National Center for Education Statistics

National Institute of Statistical Sciences
Roundtable Report

IMPUTATION IN GOVERNMENT SURVEYS

TABLE OF CONTENTS

NATIONAL INSTITUTE OF STATISTICAL SCIENCES

IMPUTATION IN GOVERNMENT SURVEYS

EXECUTIVE SUMMARY

ROUNDTABLE REPORT ON IMPUTATION IN GOVERNMENT SURVEYS
MORNING SESSION:
DISCUSSION OF HOW IMPUTATION IS USED ACROSS THE GOVERNMENT

10:15 am-10:45 am Yves Thibaudeau, PhD, U.S. Census Bureau, Center for Statistical Research and Methodology

10:45 am-11:15 am "Imputation at the BLS" - MoonJung Cho, PhD, U.S. Department of Labor, Bureau of Labor Statistics

11:15 am-11:45 am "Hitting Calibration Targets + INCA Calibration" - Cliff Spiegelman, PhD, Texas A & M University, Department of Statistics

11:45 am-12.15 pm "Three statistical issues on multiple imputation in complex survey sampling" - Jae Kwang Kim, PhD, Iowa State University, Department of Statistics

ROUNDTABLE REPORT ON IMPUTATION IN GOVERNMENT SURVEYS
AFTERNOON SESSION:
REVIEW OF ANALYSIS OF IMPUTATION APPROACHES USED IN THE NATIONAL HOUSEHOLD EDUCATION SURVEYS

1:00 pm-1:30 pm "Assessing imputation uncertainty NHES 2012" - Recai Yucel, PhD, University at Albany, SUNY, School of Public Health

1:30 pm-2:00 pm "Outline for Discussion at NCES Roundtable on Imputation" - Nat Schenker, PhD (Retired), National Center for Health Statistics, Division of Research and Methodology

2:00 pm-2:30 pm "Match Bias or Nonignorable Nonresponse? Improved Imputation and Administrative Data in the CPS ASEC" - Discussant & Moderated Discussion with Panel, Raghu Raghunathan, PhD, University of Michigan, Institute for Social Research

2:30 pm-3:00 pm Moderated Discussion between Panel and Audience - Raghu Raghunathan, PhD, University of Michigan, Institute for Social Research

3:00 pm Adjourn

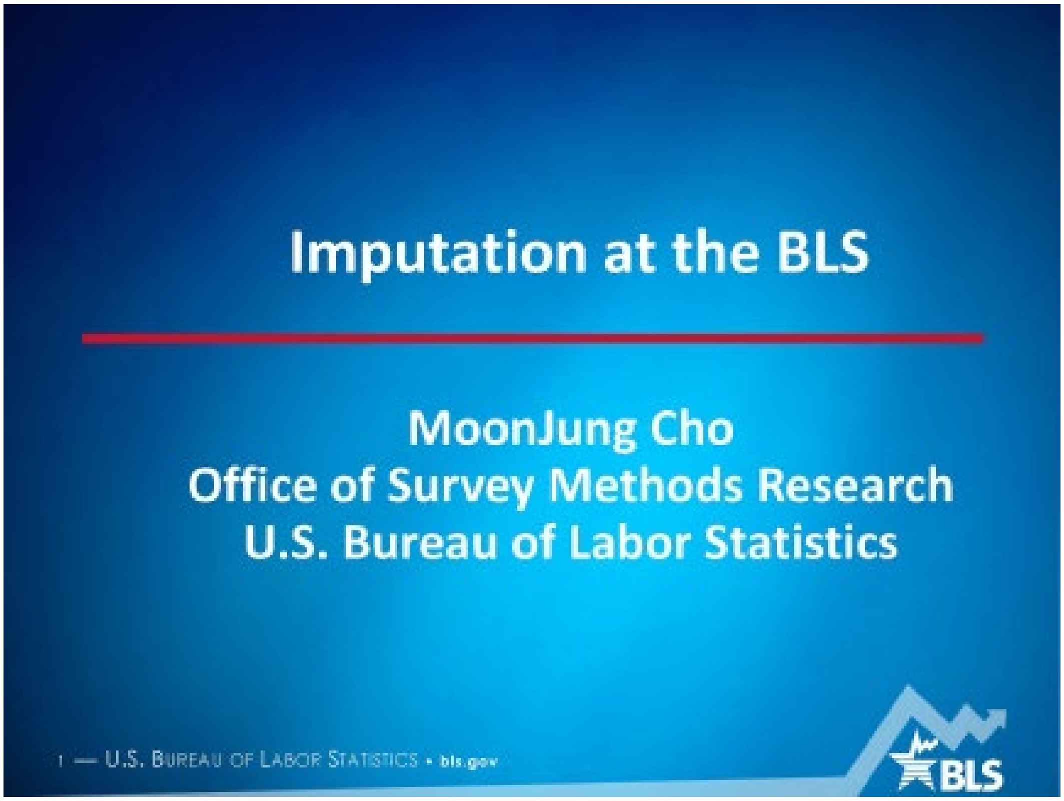
PROJECT GOAL

At the request of the National Center for Education Statistics (NCES) the National Institute of Statistical Sciences organized an Experts' Roundtable focused on the use of imputation procedures in government surveys.

On January 26, 2018, the roundtable was held in person at the NCES. Presentations showcased the practice of imputation procedures in other federal agencies such as U.S. Census Bureau, Bureau of Labor Statistics, National Agricultural Statistical Services, and National Center for Health Statistics.

Advances on imputation procedures in academic research were highlighted, with specific attention given to the potential impact of adoption of multiple imputation procedure on NCES surveys and results of a pilot study.

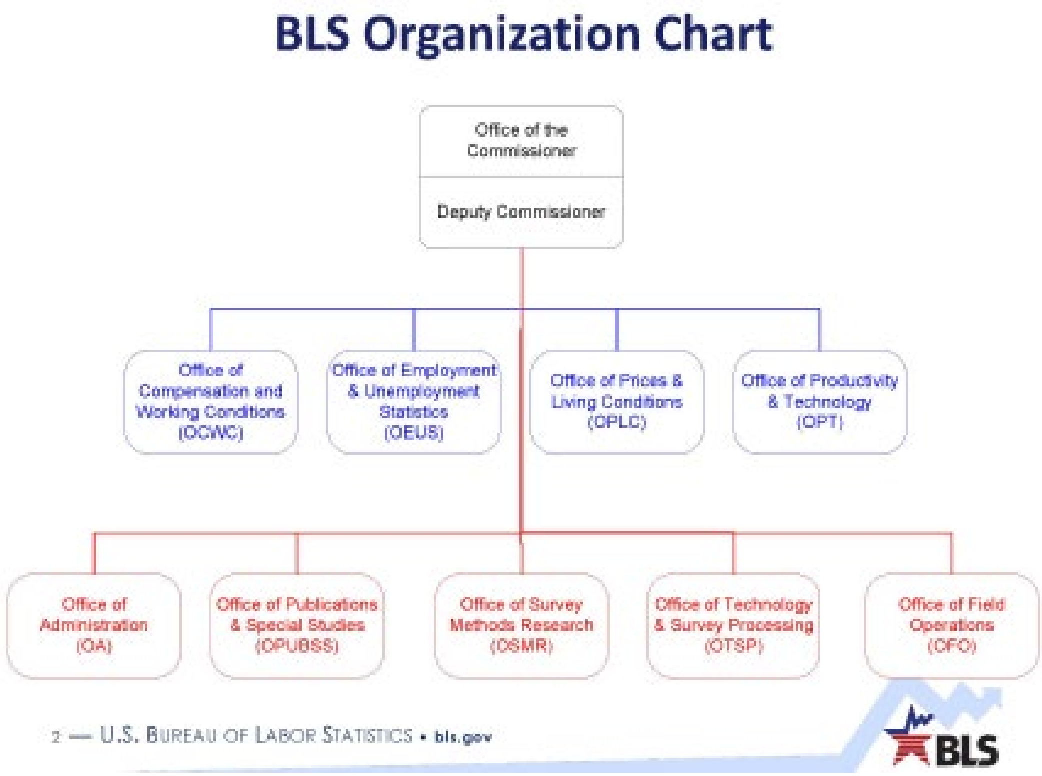
Imputation at the BLS



Imputation at the BLS

MoonJung Cho
Office of Survey Methods Research
U.S. Bureau of Labor Statistics

1 — U.S. BUREAU OF LABOR STATISTICS • bls.gov



BLS Organization Chart

Office of the Commissioner

Deputy Commissioner

Office of Compensation and Working Conditions (OCWC)

Office of Employment & Unemployment Statistics (OEUS)

Office of Prices & Living Conditions (OPLC)

Office of Productivity & Technology (OPT)

Office of Administration (OA)

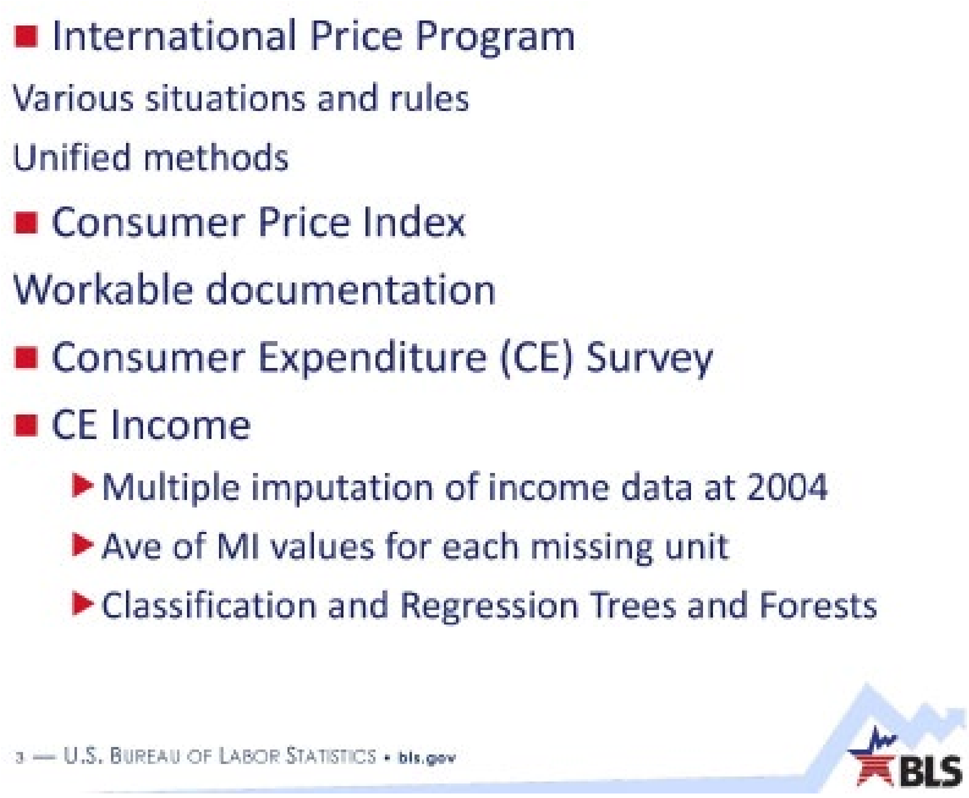
Office of Publications & Special Studies (OPUBSS)

Office of Survey Methods Research (OSMR)

Office of Technology & Survey Processing (OTSP)

Office of Field Operations (OFO)

2 — U.S. BUREAU OF LABOR STATISTICS • bls.gov



- International Price Program

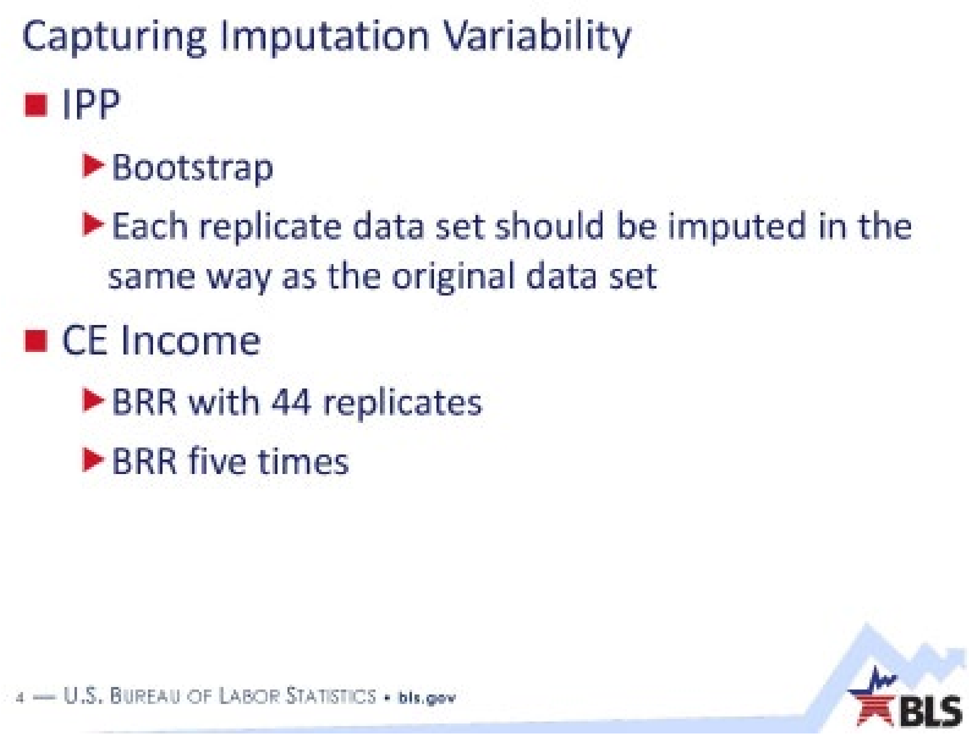
Various situations and rules

Unified methods

- Consumer Price Index

Workable documentation

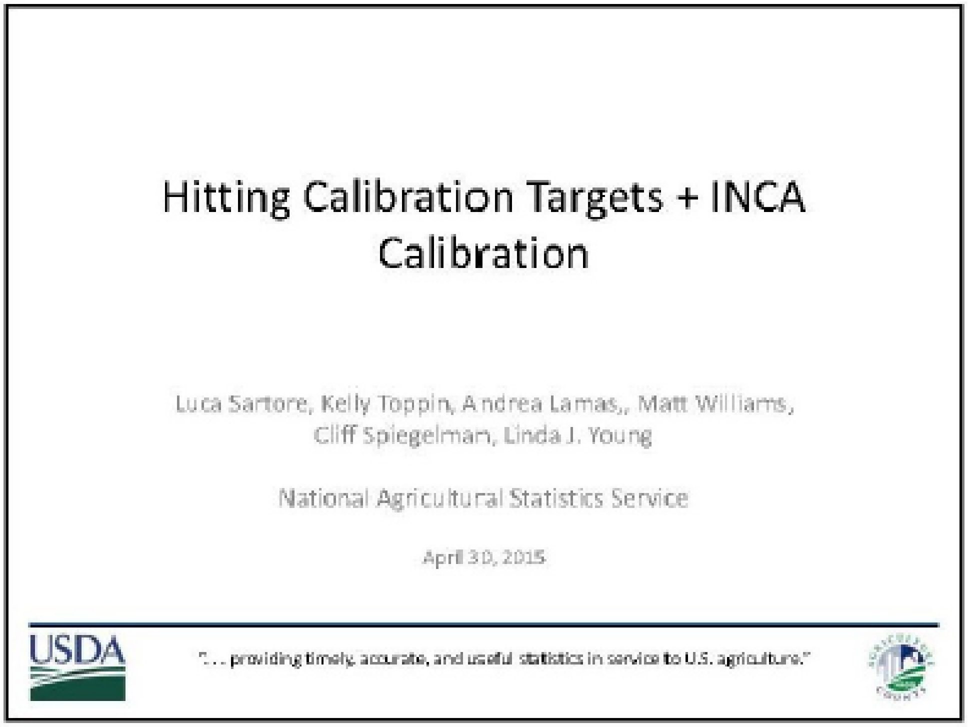
- Consumer Expenditure (CE) Survey
- CE Income
  - Multiple imputation of income data at 2004
  - Ave of MI values for each missing unit
  - Classification and Regression Trees and Forests



Capturing Imputation Variability

- IPP
  - Bootstrap
  - Each replicate data set should be imputed in the same way as the original data set
- CE Income
  - BRR with 44 replicates
  - BRR five times

Hitting Calibration Targets + INCA Calibration



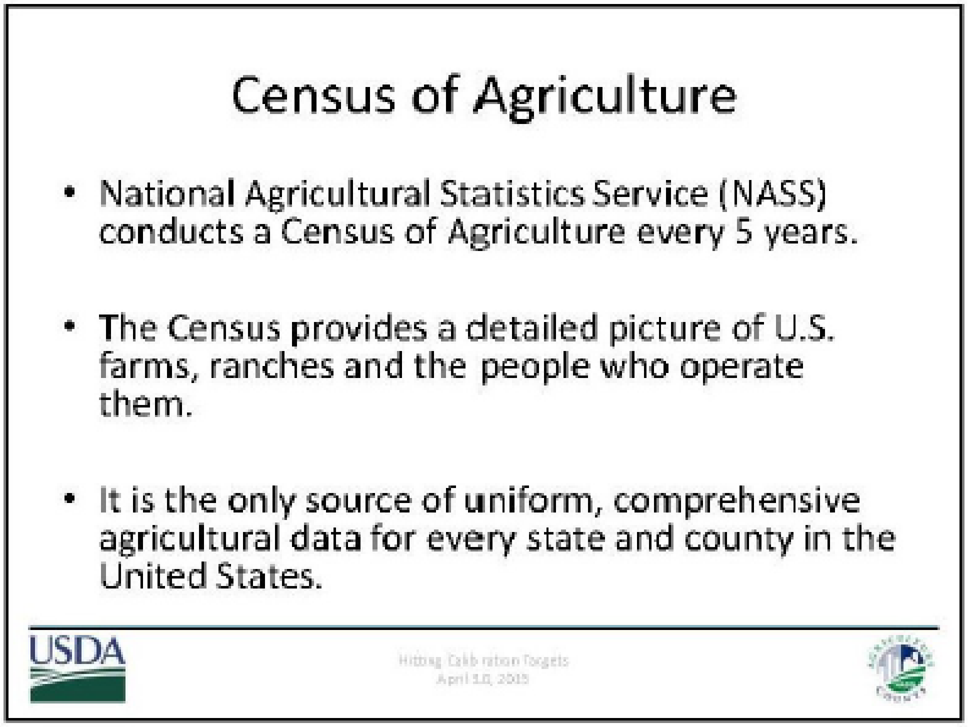
Hitting Calibration Targets + INCA Calibration

Luca Sartore, Kelly Toppin, Andrea Lamas,, Matt Williams,
Cliff Spiegelman, Linda J. Young

National Agricultural Statistics Service

April 30, 2015

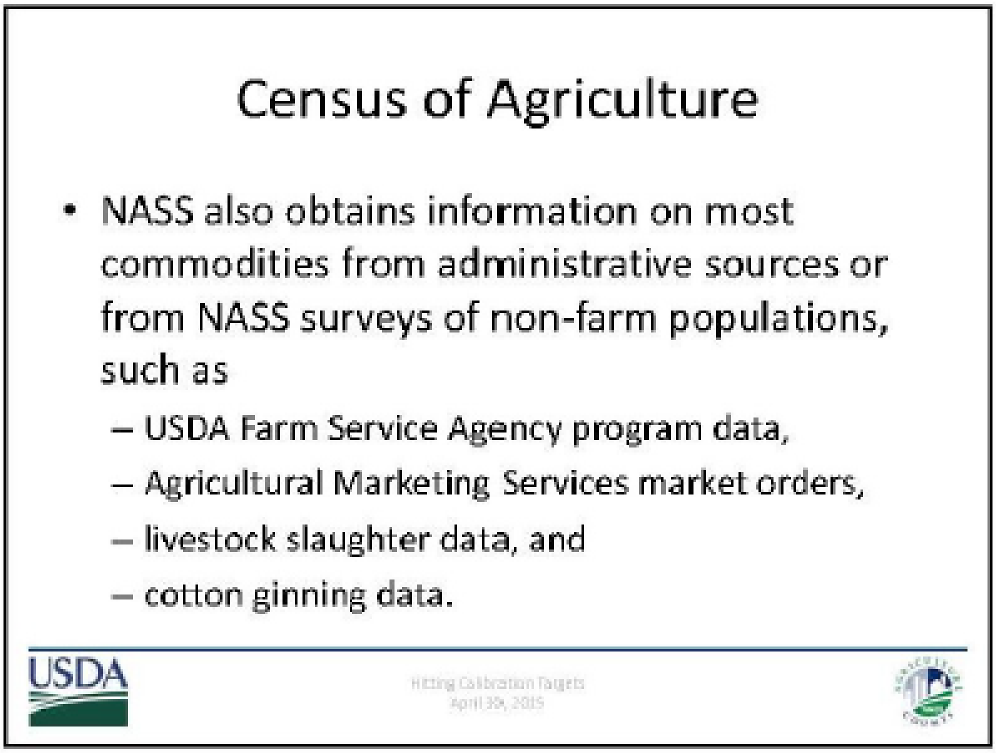
"... providing timely, accurate, and useful statistics in service to U.S. agriculture."



Census of Agriculture

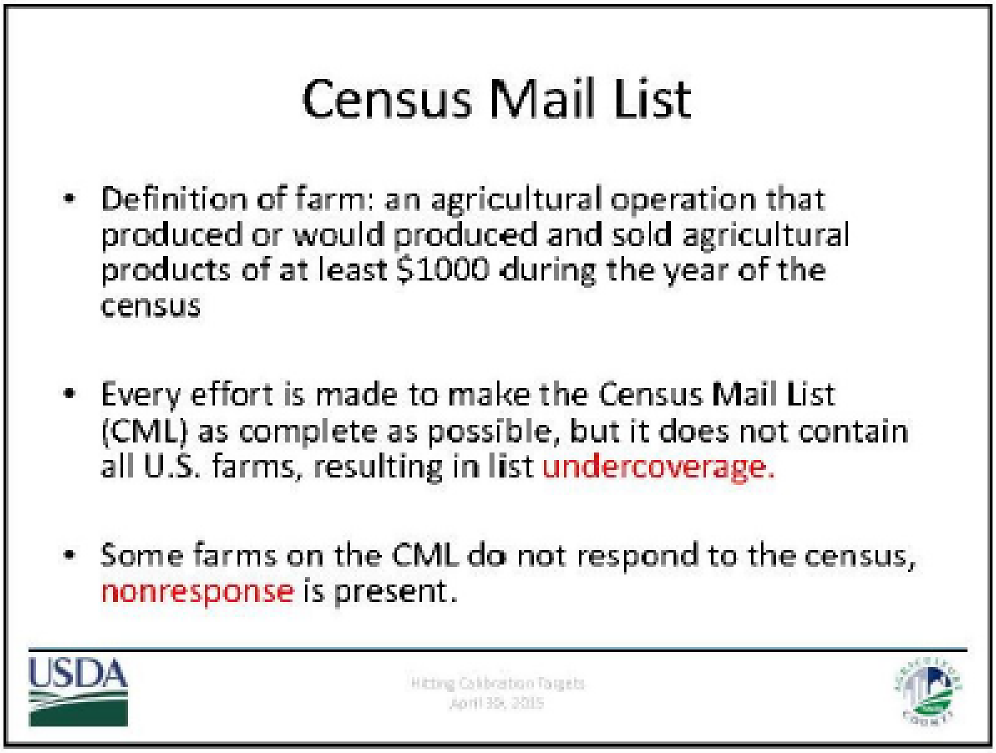
- National Agricultural Statistics Service (NASS) conducts a Census of Agriculture every 5 years.
- The Census provides a detailed picture of U.S. farms, ranches and the people who operate them.
- It is the only source of uniform, comprehensive agricultural data for every state and county in the United States.

Hitting Calibration Targets
April 30, 2015



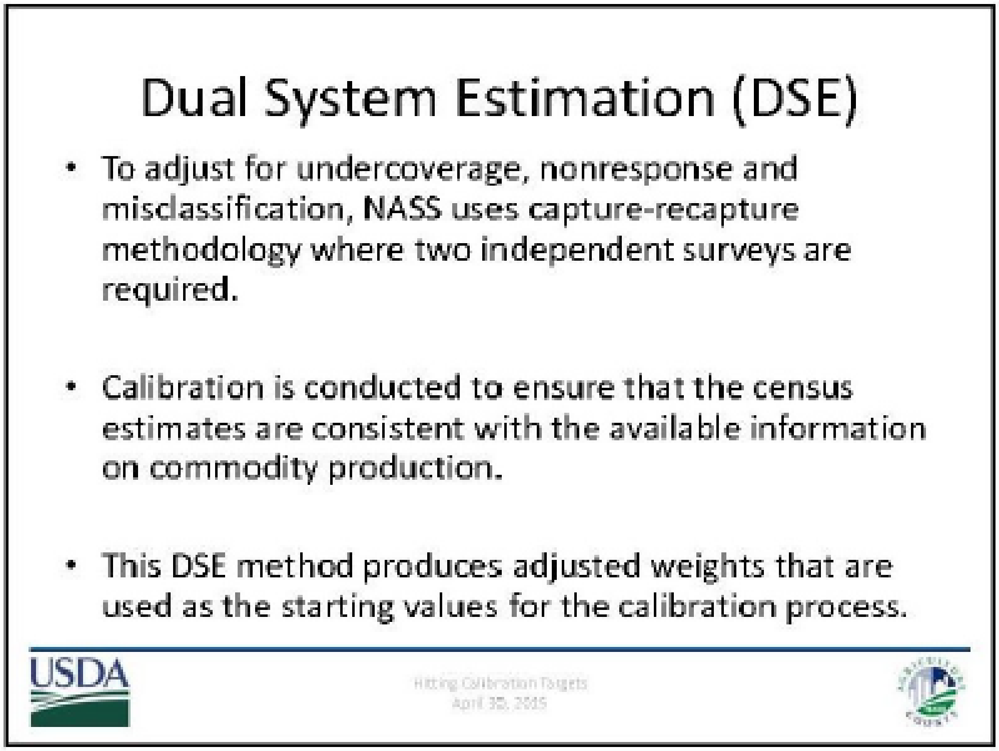
Census of Agriculture

- NASS also obtains information on most commodities from administrative sources or from NASS surveys of non-farm populations, such as
  - USDA Farm Service Agency program data,
  - Agricultural Marketing Services market orders,
  - livestock slaughter data, and
  - cotton ginning data.



Census Mail List

- Definition of farm: an agricultural operation that produced or would produced and sold agricultural products of at least $1000 during the year of the census
- Every effort is made to make the Census Mail List (CML) as complete as possible, but it does not contain all U.S. farms, resulting in list undercoverage.
- Some farms on the CML do not respond to the census, nonresponse is present.



Dual System Estimation (DSE)

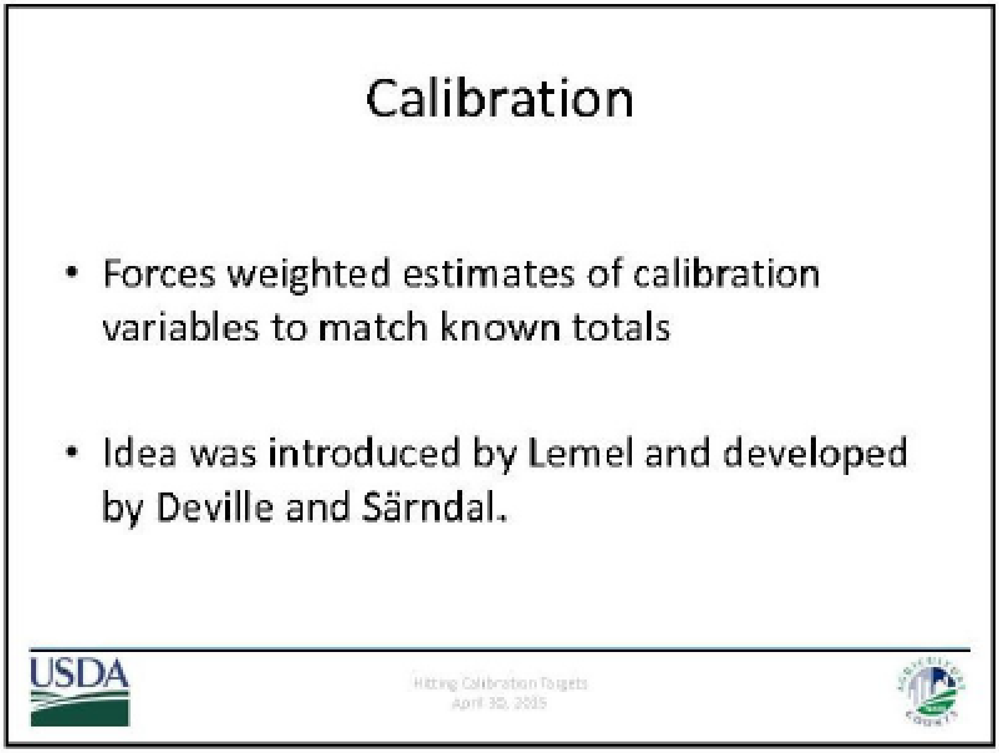
- To adjust for undercoverage, nonresponse and misclassification, NASS uses capture-recapture methodology where two independent surveys are required.

- Calibration is conducted to ensure that the census estimates are consistent with the available information on commodity production.

- This DSE method produces adjusted weights that are used as the starting values for the calibration process.

USDA

Hitting Calibration Targets
April 30, 2015



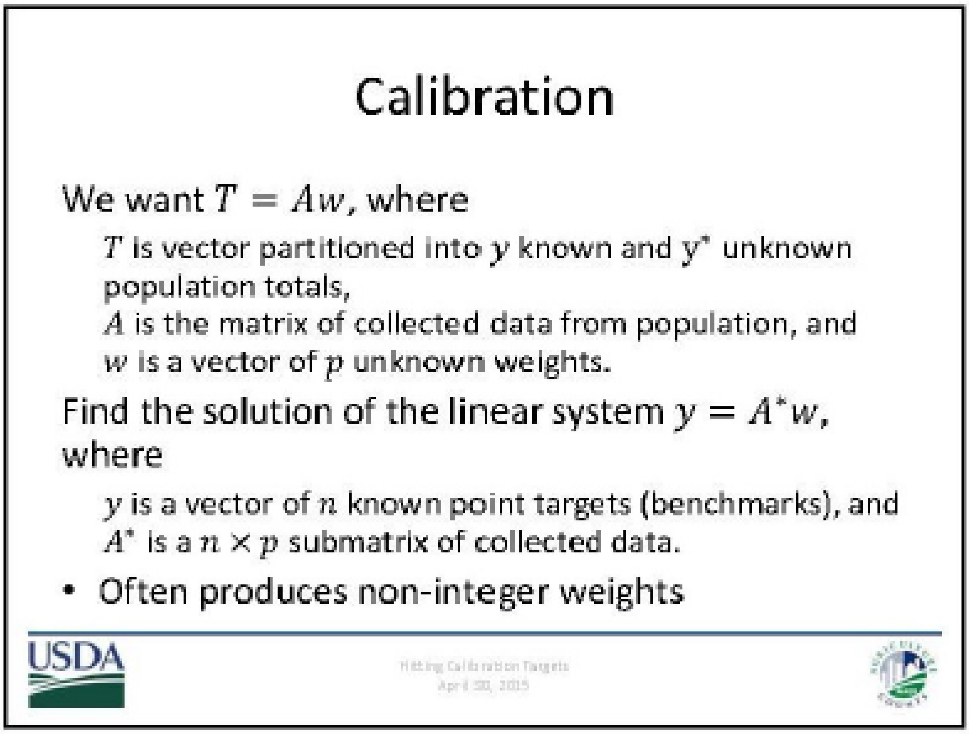
Calibration

- Forces weighted estimates of calibration variables to match known totals

- Idea was introduced by Lemel and developed by Deville and Särndal.

USDA

Hitting Calibration Targets
April 30, 2015



Calibration

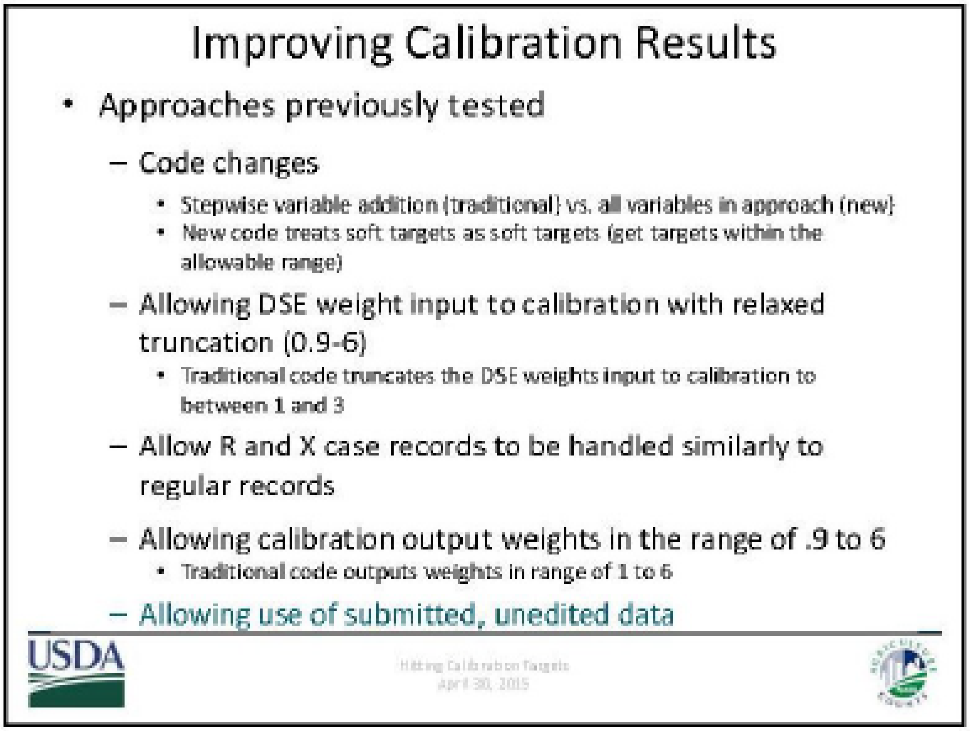
We want $T = Aw$, where

T is vector partitioned into y known and y^* unknown population totals,
A is the matrix of collected data from population, and
w is a vector of p unknown weights.

Find the solution of the linear system $y = A^*w$, where

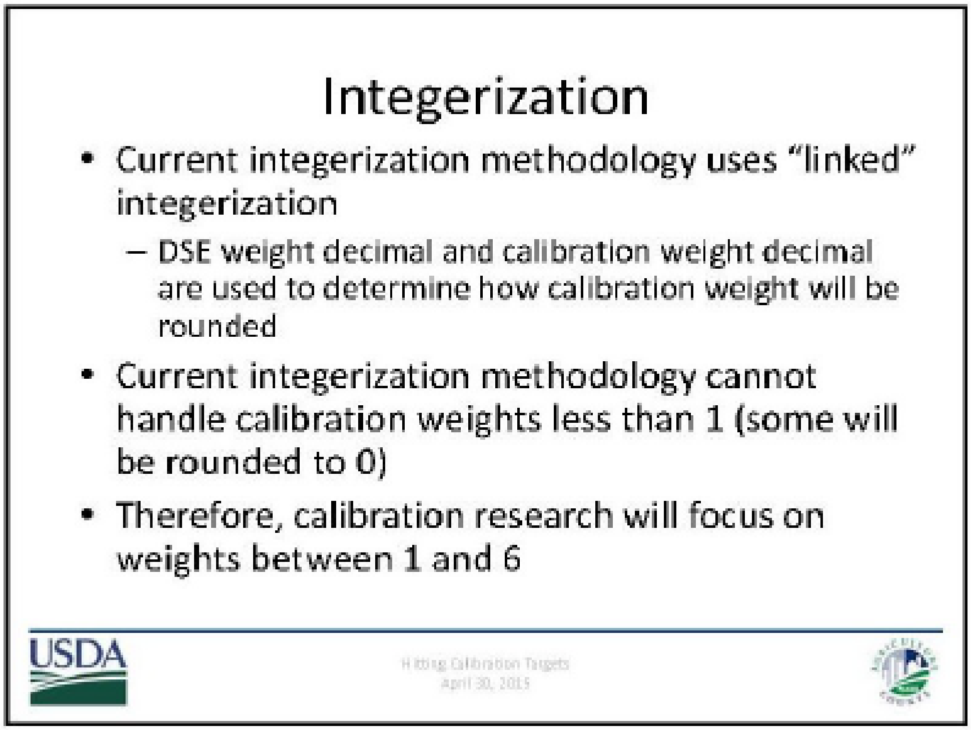
y is a vector of n known point targets (benchmarks), and
A^* is a $n \times p$ submatrix of collected data.

- Often produces non-integer weights



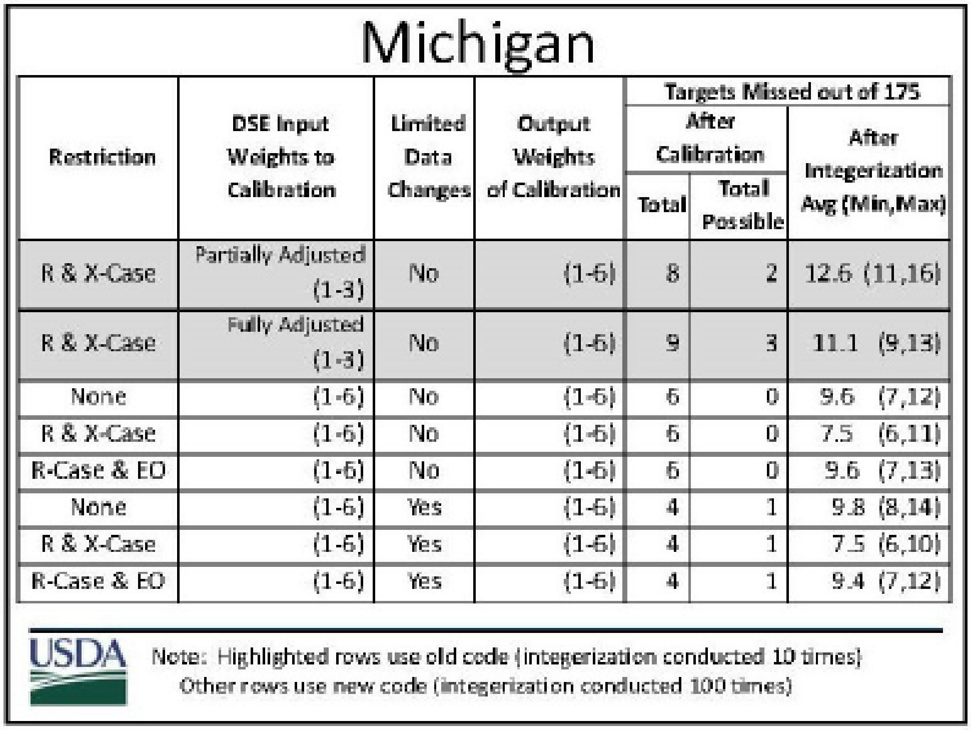
Improving Calibration Results

- Approaches previously tested
  - Code changes
    - Stepwise variable addition (traditional) vs. all variables in approach (new)
    - New code treats soft targets as soft targets (get targets within the allowable range)
  - Allowing DSE weight input to calibration with relaxed truncation (0.9-6)
    - Traditional code truncates the DSE weights input to calibration to between 1 and 3
  - Allow R and X case records to be handled similarly to regular records
  - Allowing calibration output weights in the range of .9 to 6
    - Traditional code outputs weights in range of 1 to 6
  - Allowing use of submitted, unedited data



Integerization

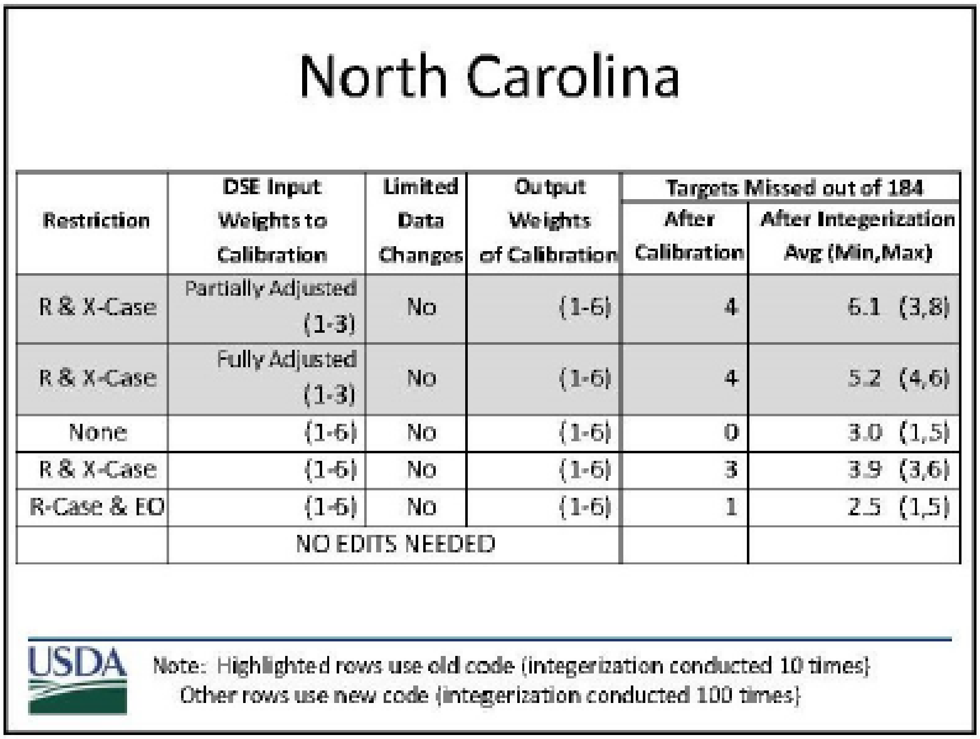
- Current integerization methodology uses “linked” integerization
  - DSE weight decimal and calibration weight decimal are used to determine how calibration weight will be rounded
- Current integerization methodology cannot handle calibration weights less than 1 (some will be rounded to 0)
- Therefore, calibration research will focus on weights between 1 and 6



Michigan

Restriction	DSE Input Weights to Calibration	Limited Data Changes	Output Weights of Calibration	Targets Missed out of 175		
				After Calibration		After Integerization Avg (Min,Max)
				Total	Total Possible	
R & X-Case	Partially Adjusted (1-3)	No	(1-6)	8	2	12.6 (11,16)
R & X-Case	Fully Adjusted (1-3)	No	(1-6)	9	3	11.1 (9,13)
None	(1-6)	No	(1-6)	6	0	9.6 (7,12)
R & X-Case	(1-6)	No	(1-6)	6	0	7.5 (6,11)
R-Case & EO	(1-6)	No	(1-6)	6	0	9.6 (7,13)
None	(1-6)	Yes	(1-6)	4	1	9.8 (8,14)
R & X-Case	(1-6)	Yes	(1-6)	4	1	7.5 (6,10)
R-Case & EO	(1-6)	Yes	(1-6)	4	1	9.4 (7,12)

Note: Highlighted rows use old code (integerization conducted 10 times)
Other rows use new code (integerization conducted 100 times)

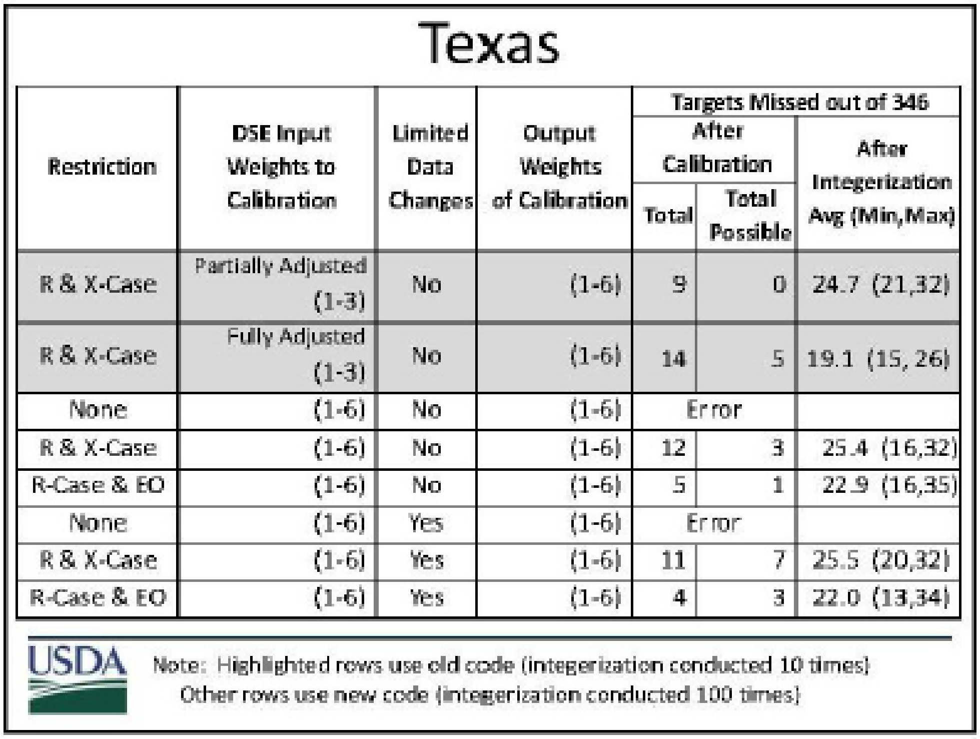


North Carolina

Restriction	DSE Input Weights to Calibration	Limited Data Changes	Output Weights of Calibration	Targets Missed out of 184	
				After Calibration	After Integerization Avg (Min,Max)
R & X-Case	Partially Adjusted (1-3)	No	(1-6)	4	6.1 (3,8)
R & X-Case	Fully Adjusted (1-3)	No	(1-6)	4	5.2 (4,6)
None	(1-6)	No	(1-6)	0	3.0 (1,5)
R & X-Case	(1-6)	No	(1-6)	3	3.9 (3,6)
R-Case & EO	(1-6)	No	(1-6)	1	2.5 (1,5)
	NO EDITS NEEDED				

USDA

Note: Highlighted rows use old code (integerization conducted 10 times)
Other rows use new code (integerization conducted 100 times)

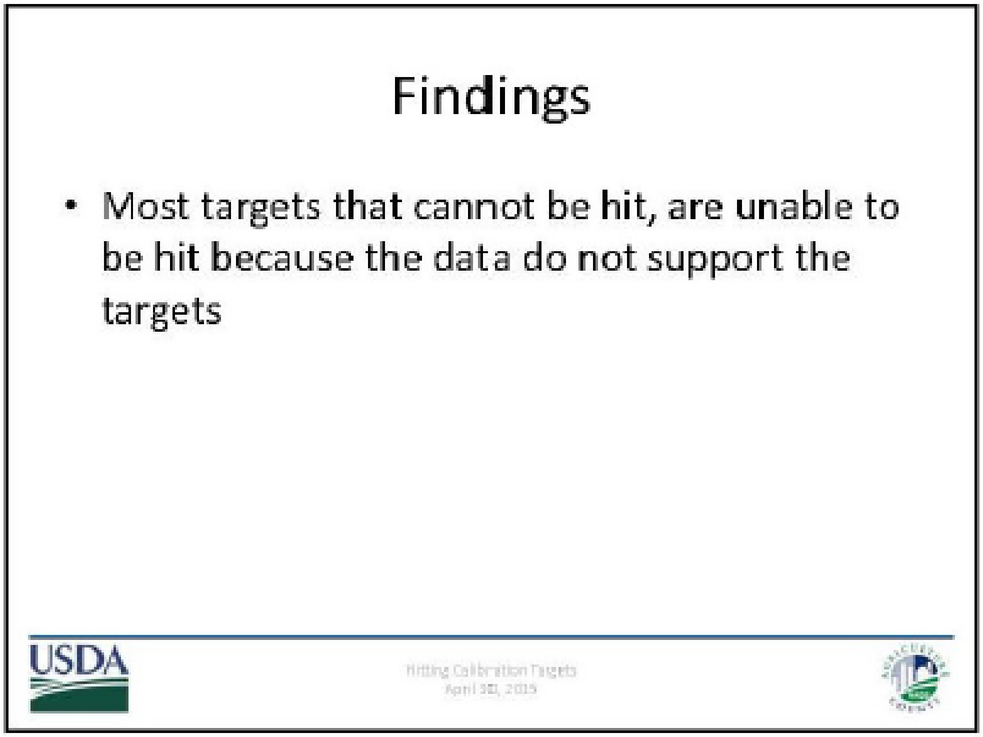


Texas

Restriction	DSE Input Weights to Calibration	Limited Data Changes	Output Weights of Calibration	Targets Missed out of 346		
				After Calibration		After Integerization Avg (Min,Max)
				Total	Total Possible	
R & X-Case	Partially Adjusted (1-3)	No	(1-6)	9	0	24.7 (21,32)
R & X-Case	Fully Adjusted (1-3)	No	(1-6)	14	5	19.1 (15, 26)
None	(1-6)	No	(1-6)	Error		
R & X-Case	(1-6)	No	(1-6)	12	3	25.4 (16,32)
R-Case & EO	(1-6)	No	(1-6)	5	1	22.9 (16,35)
None	(1-6)	Yes	(1-6)	Error		
R & X-Case	(1-6)	Yes	(1-6)	11	7	25.5 (20,32)
R-Case & EO	(1-6)	Yes	(1-6)	4	3	22.0 (13,34)

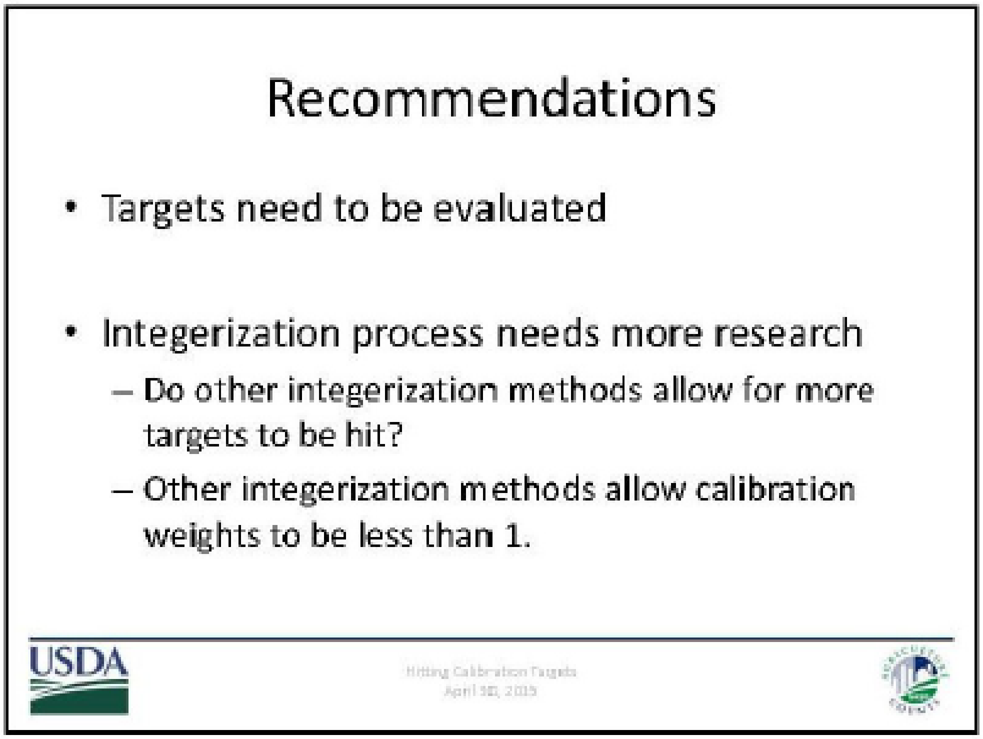
USDA

Note: Highlighted rows use old code (integerization conducted 10 times)
Other rows use new code (integerization conducted 100 times)



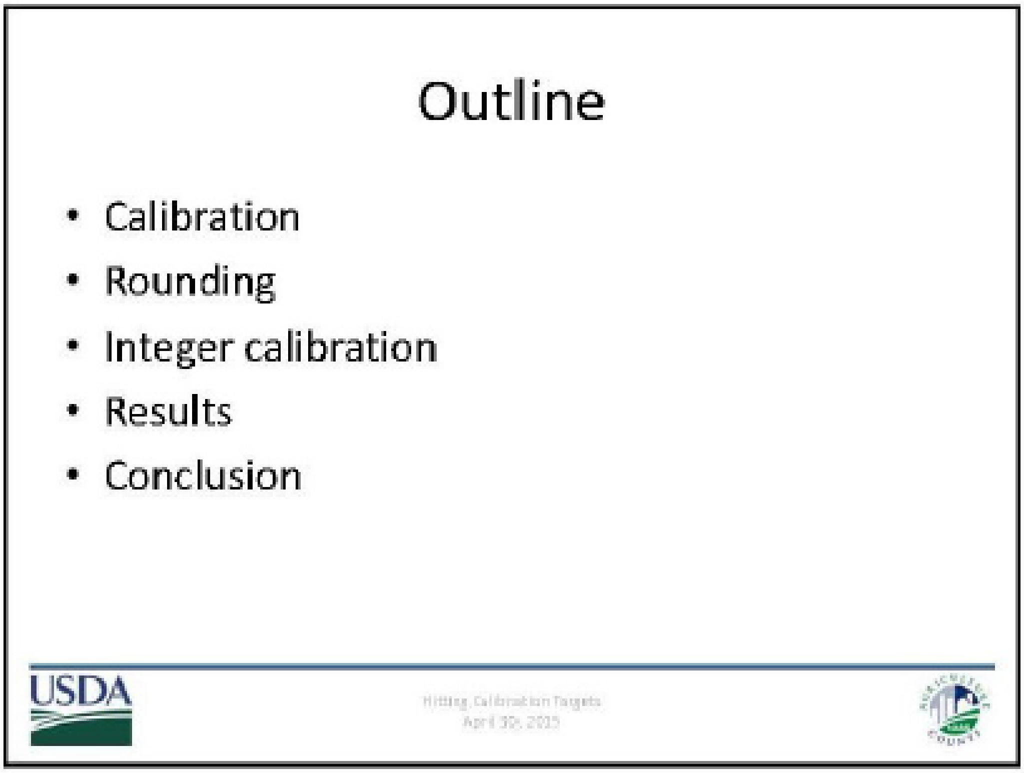
Findings

- Most targets that cannot be hit, are unable to be hit because the data do not support the targets



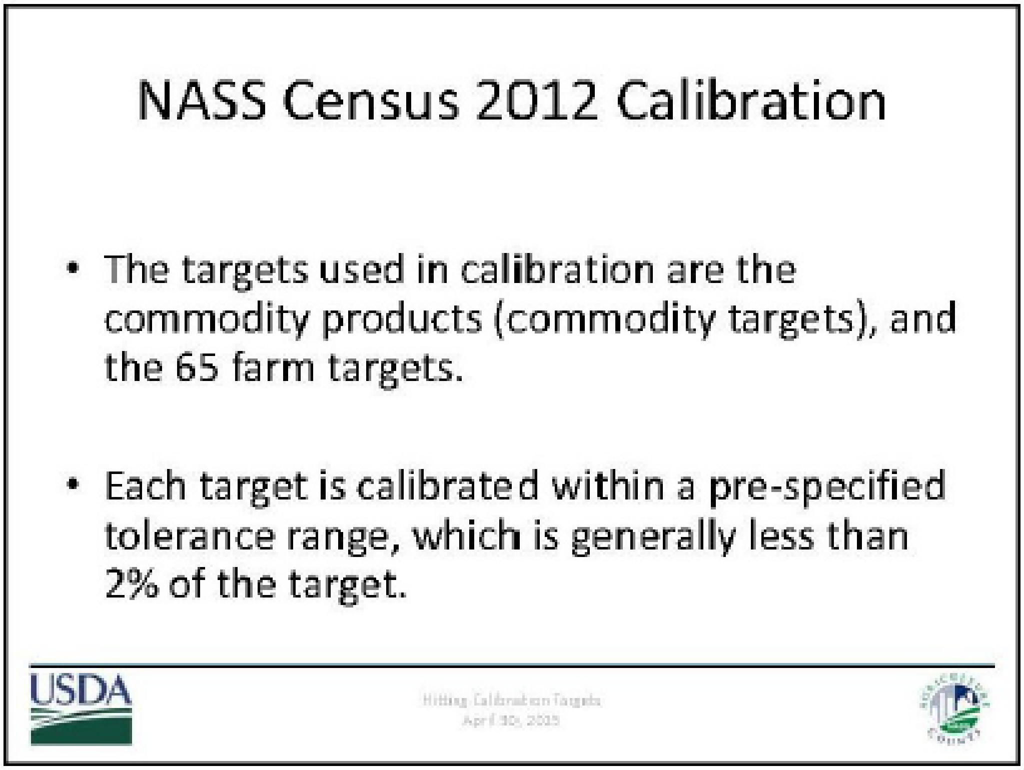
Recommendations

- Targets need to be evaluated
- Integerization process needs more research
  - Do other integerization methods allow for more targets to be hit?
  - Other integerization methods allow calibration weights to be less than 1.



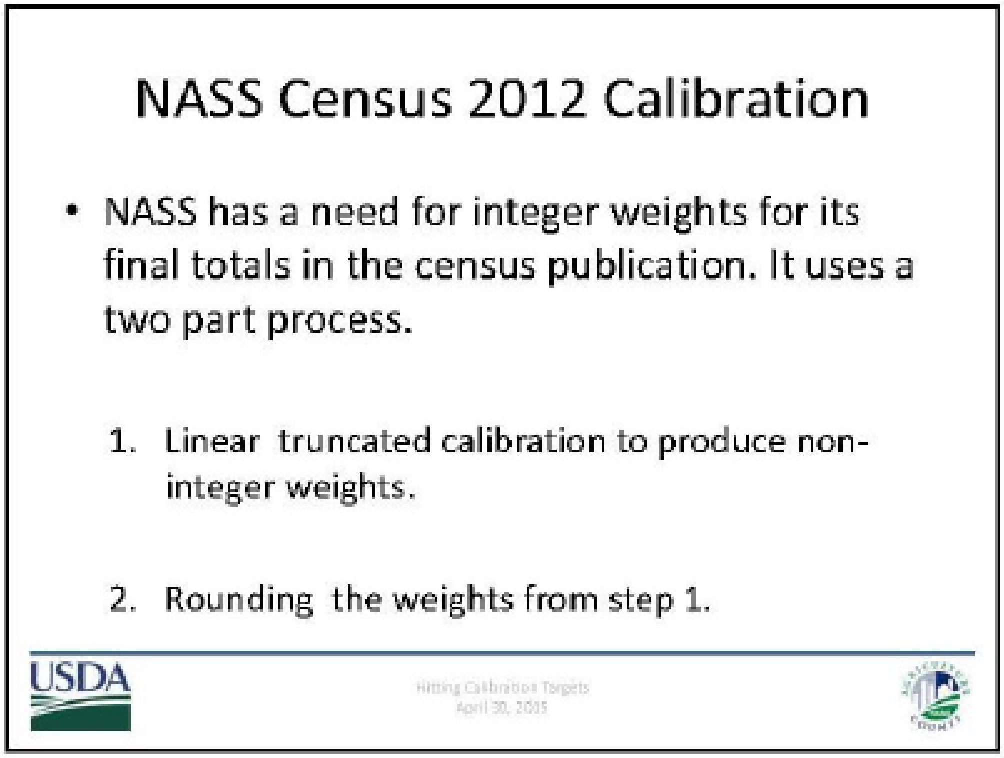
Outline

- Calibration
- Rounding
- Integer calibration
- Results
- Conclusion



NASS Census 2012 Calibration

- The targets used in calibration are the commodity products (commodity targets), and the 65 farm targets.
- Each target is calibrated within a pre-specified tolerance range, which is generally less than 2% of the target.



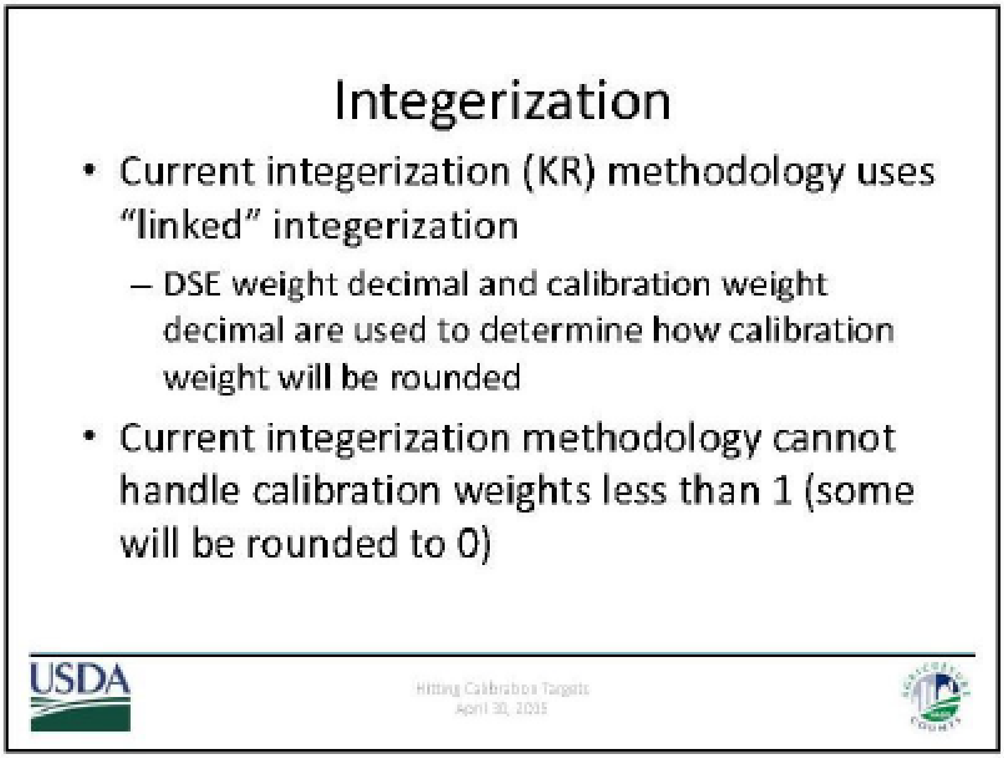
NASS Census 2012 Calibration

- NASS has a need for integer weights for its final totals in the census publication. It uses a two part process.

1. Linear truncated calibration to produce non-integer weights.

2. Rounding the weights from step 1.

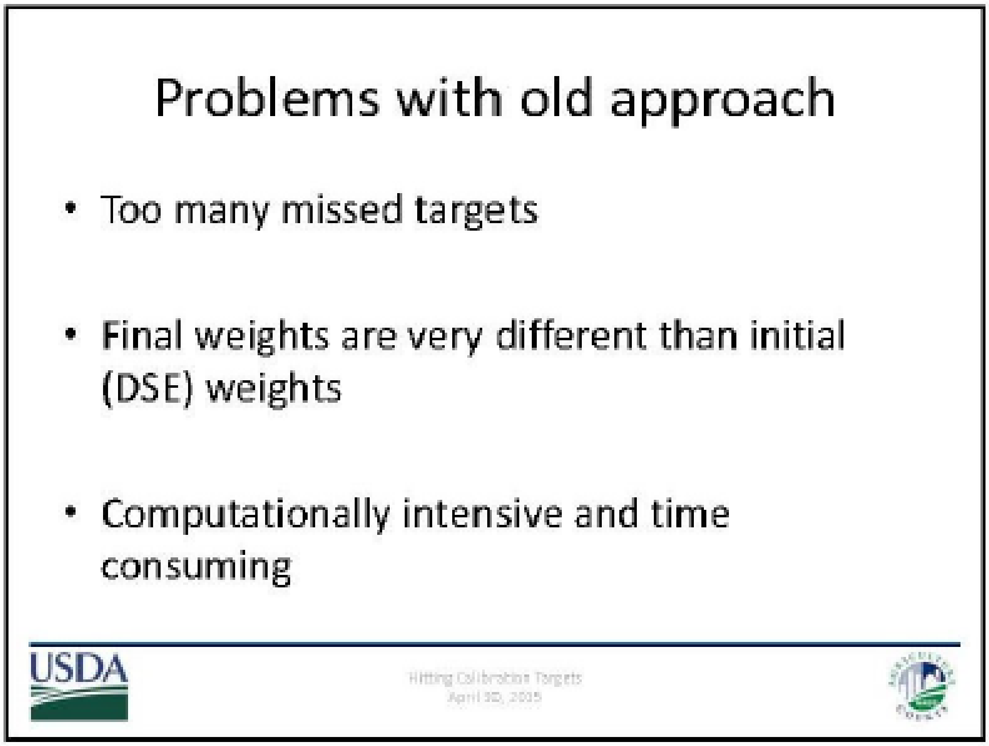
Hitting Calibration Targets
April 30, 2015



Integerization

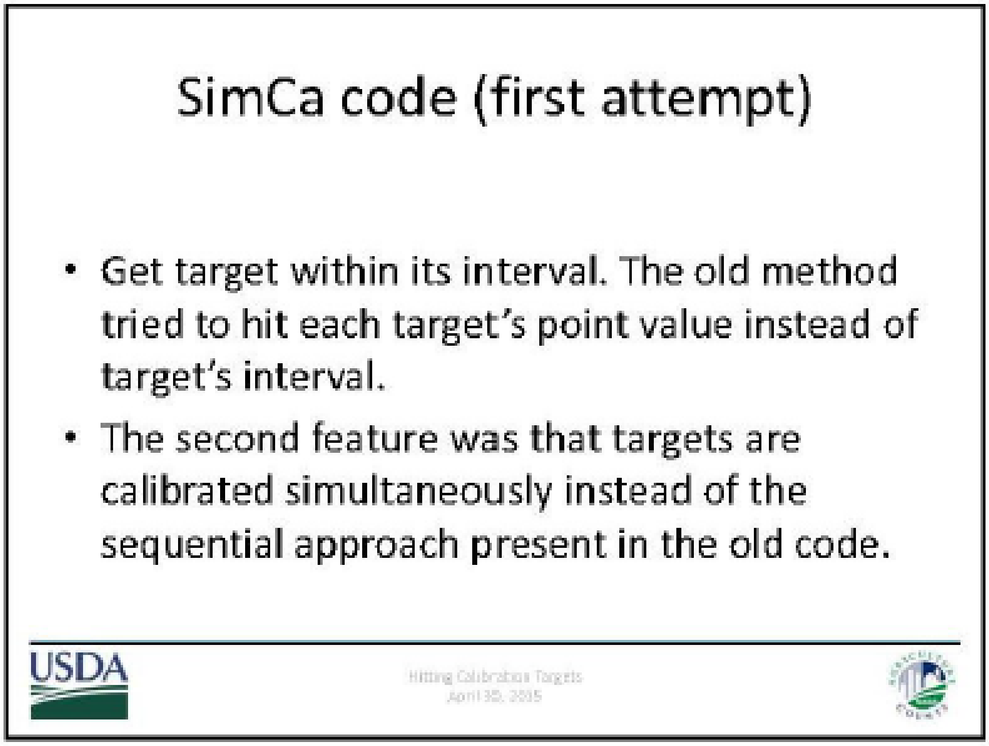
- Current integerization (KR) methodology uses “linked” integerization
  - DSE weight decimal and calibration weight decimal are used to determine how calibration weight will be rounded
- Current integerization methodology cannot handle calibration weights less than 1 (some will be rounded to 0)

Hitting Calibration Targets
April 30, 2015



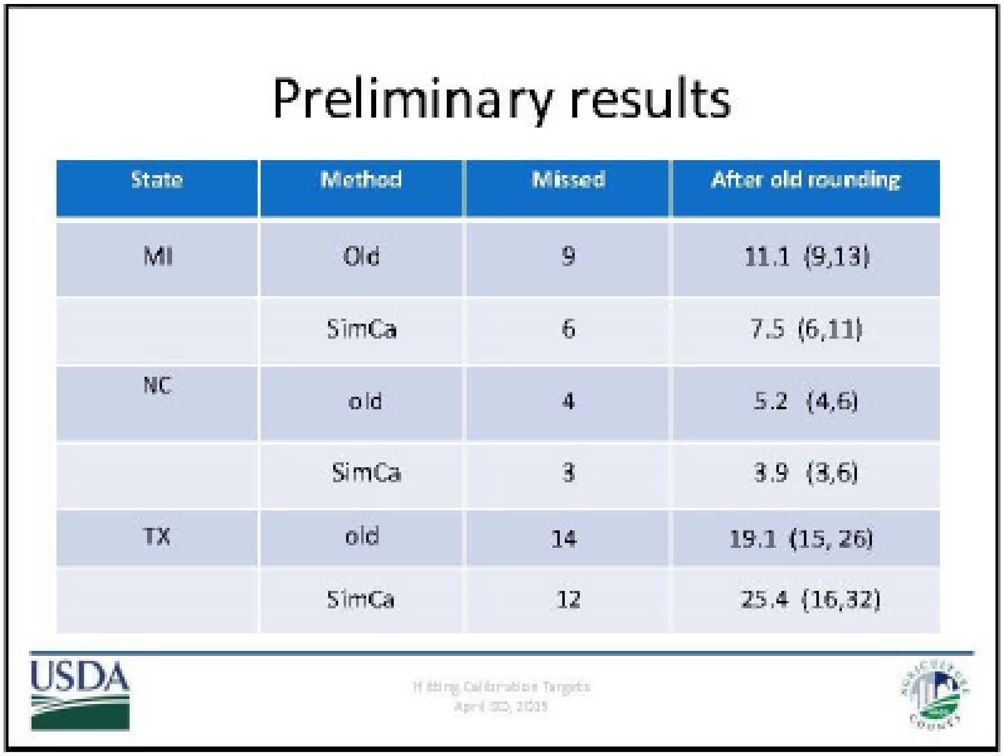
Problems with old approach

- Too many missed targets
- Final weights are very different than initial (DSE) weights
- Computationally intensive and time consuming



SimCa code (first attempt)

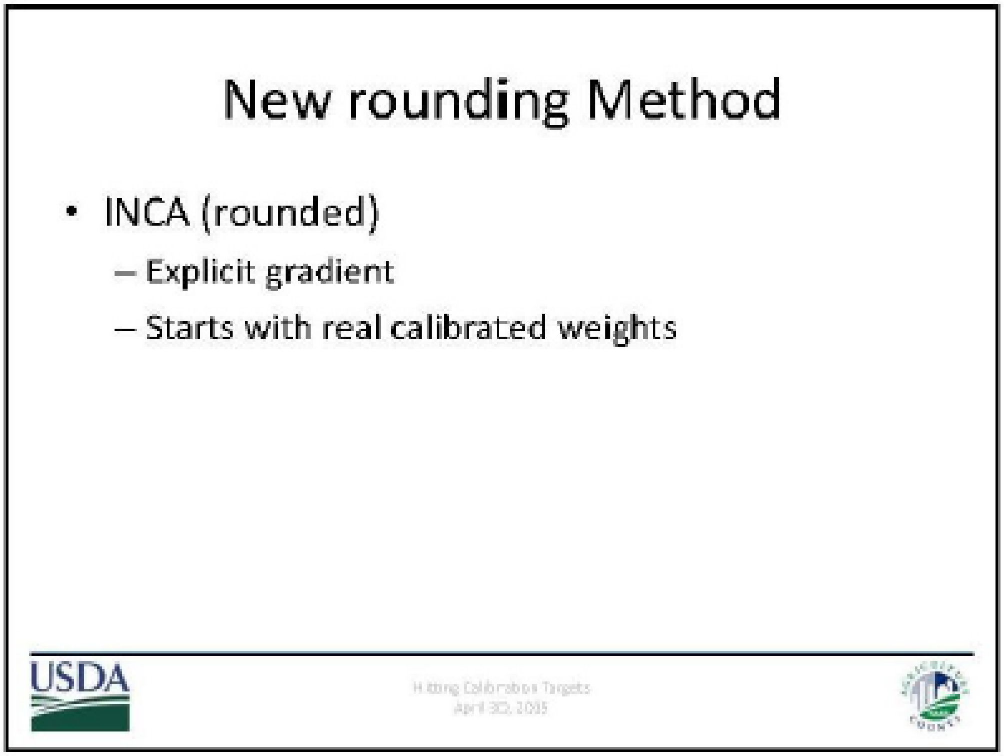
- Get target within its interval. The old method tried to hit each target's point value instead of target's interval.
- The second feature was that targets are calibrated simultaneously instead of the sequential approach present in the old code.



Preliminary results

State	Method	Missed	After old rounding
MI	Old	9	11.1 (9,13)
	SimCa	6	7.5 (6,11)
NC	old	4	5.2 (4,6)
	SimCa	3	3.9 (3,6)
TX	old	14	19.1 (15, 26)
	SimCa	12	25.4 (16,32)

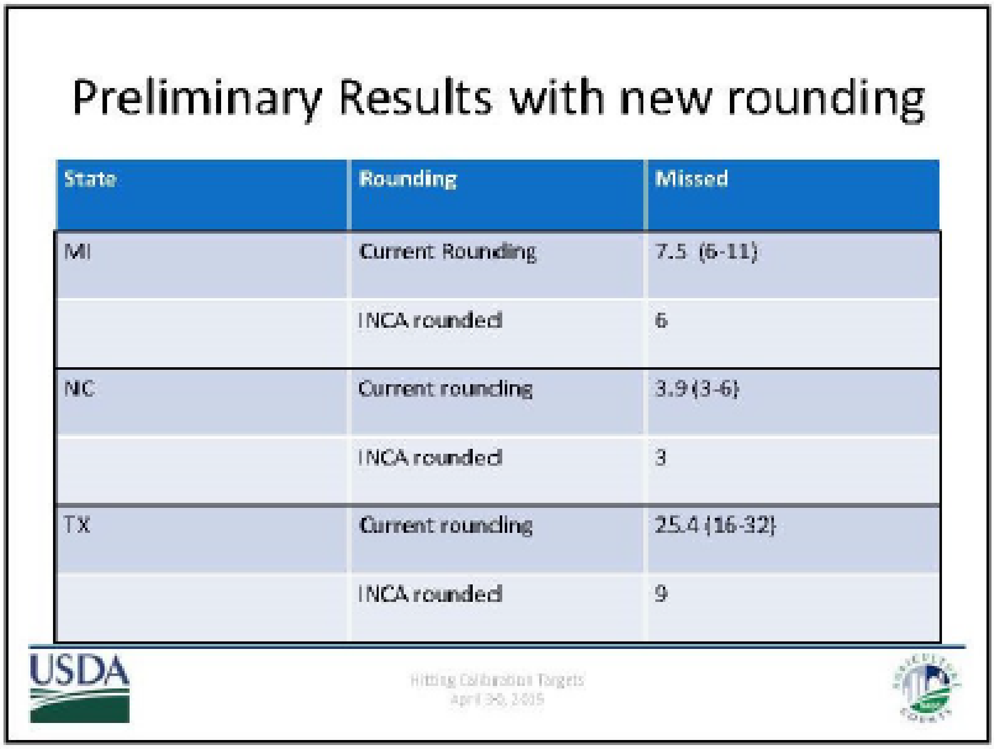
Hitting Calibration Targets
April 30, 2015



New rounding Method

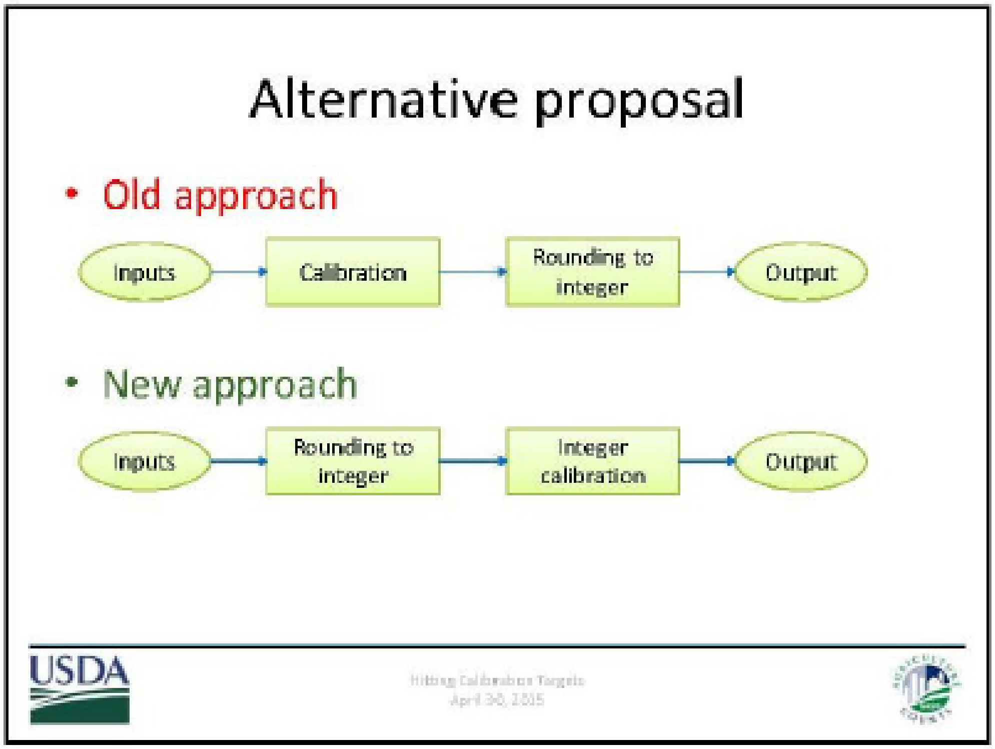
- INCA (rounded)
  - Explicit gradient
  - Starts with real calibrated weights

Hitting Calibration Targets
April 30, 2015



Preliminary Results with new rounding

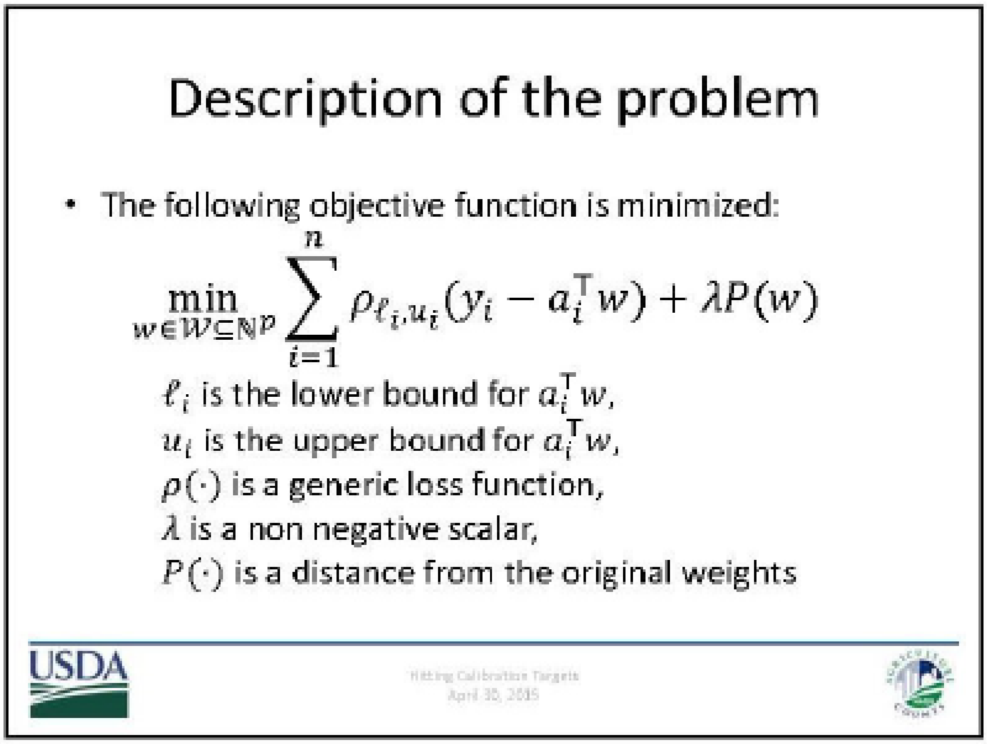
State	Rounding	Missed
MI	Current Rounding	7.5 (6-11)
	INCA rounded	6
NC	Current rounding	3.9 (3-6)
	INCA rounded	3
TX	Current rounding	25.4 (16-32)
	INCA rounded	9



Alternative proposal

- Old approach

- New approach



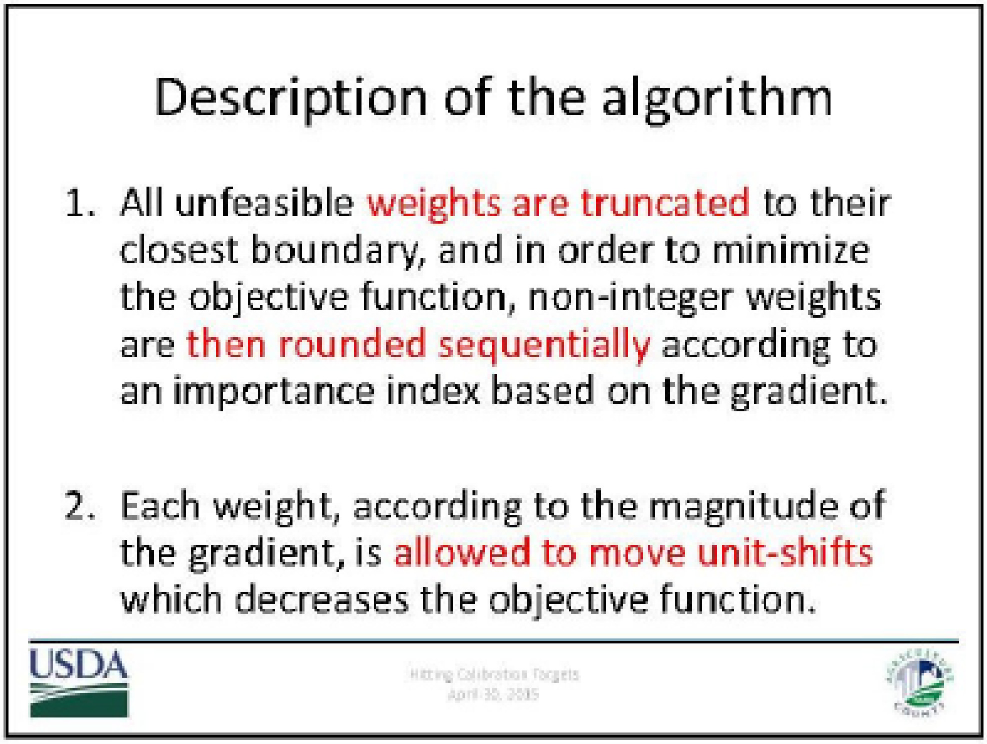
Description of the problem

- The following objective function is minimized:

$$\min_{w\in\mathcal{W}\subseteq\mathbb{N}^p} \sum_{i=1}^{n} \rho_{\ell_i,u_i}(y_i - a_i^\top w) + \lambda P(w)$$

ℓ_i is the lower bound for $a_i^\top w$,
u_i is the upper bound for $a_i^\top w$,
$\rho(\cdot)$ is a generic loss function,
λ is a non negative scalar,
$P(\cdot)$ is a distance from the original weights

Hitting Calibration Targets
April 30, 2015

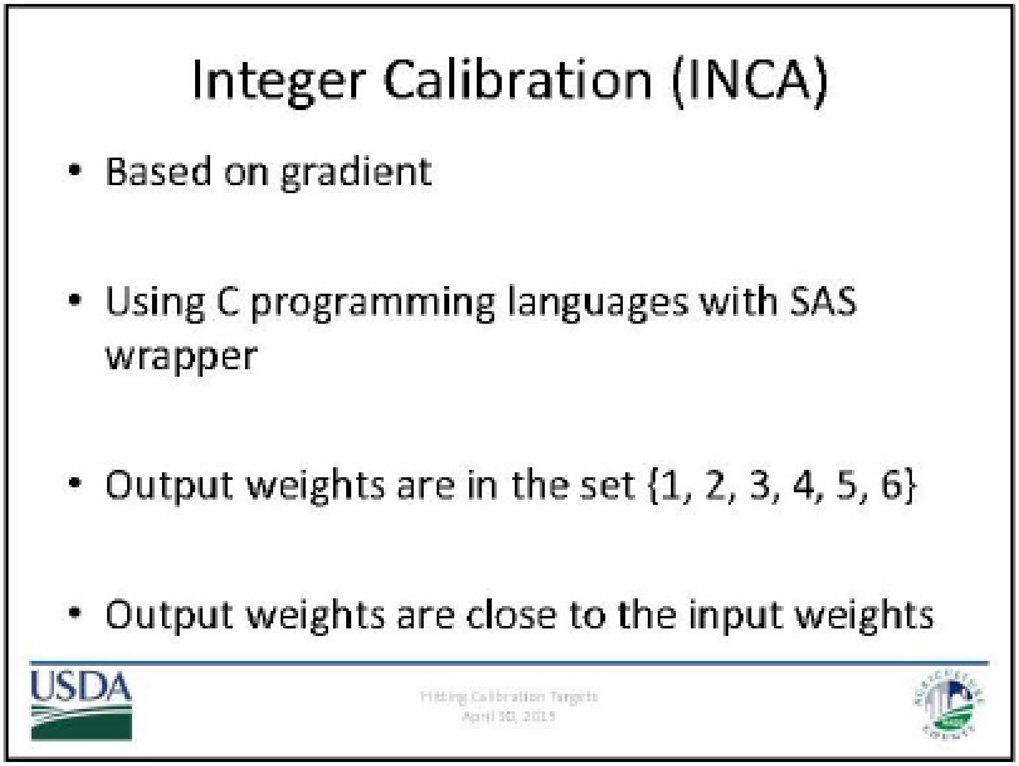


Description of the algorithm

1. All unfeasible weights are truncated to their closest boundary, and in order to minimize the objective function, non-integer weights are then rounded sequentially according to an importance index based on the gradient.

2. Each weight, according to the magnitude of the gradient, is allowed to move unit-shifts which decreases the objective function.

Hitting Calibration Targets
April 30, 2015

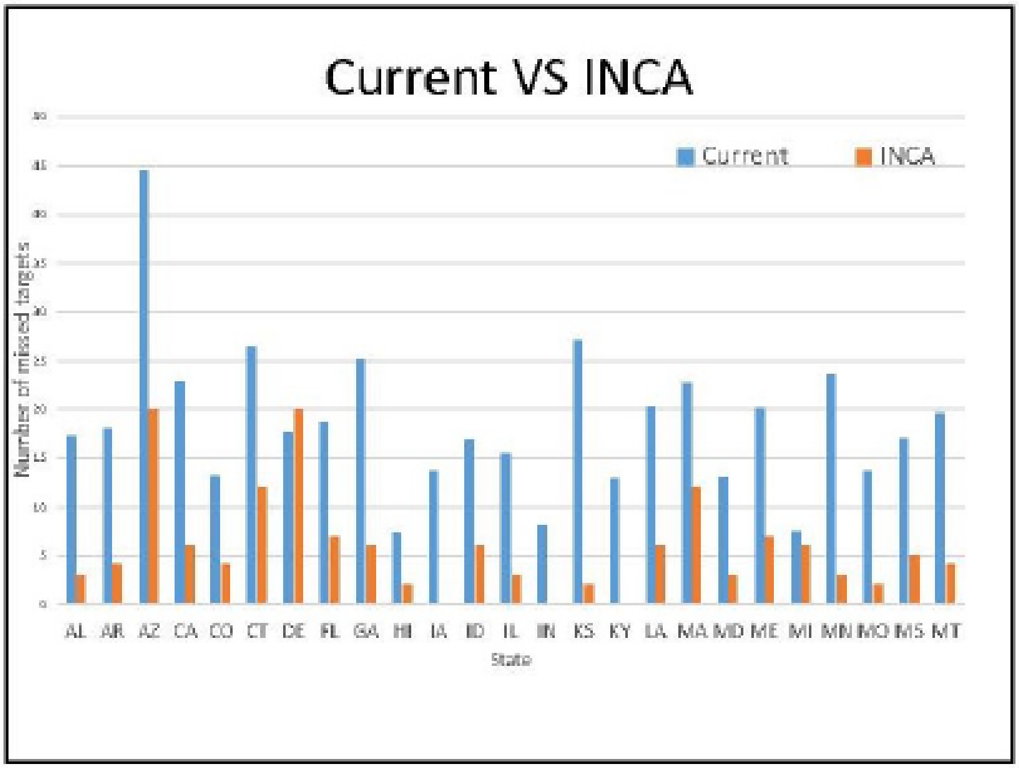


Integer Calibration (INCA)

- Based on gradient
- Using C programming languages with SAS wrapper
- Output weights are in the set {1, 2, 3, 4, 5, 6}
- Output weights are close to the input weights

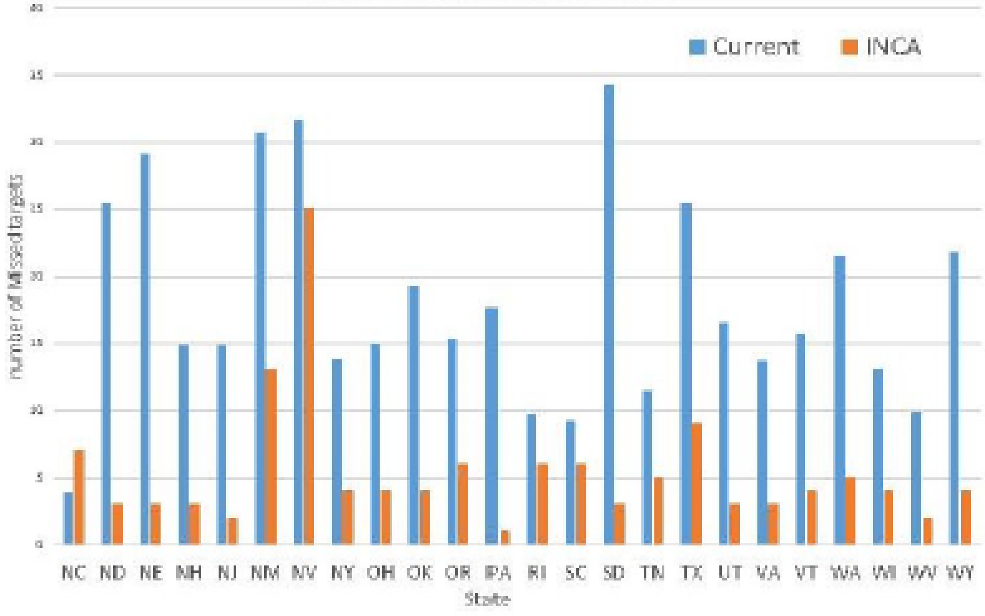
USDA

Hitting Calibration Targets
April 30, 2015



Current VS INCA

Current VS INCA

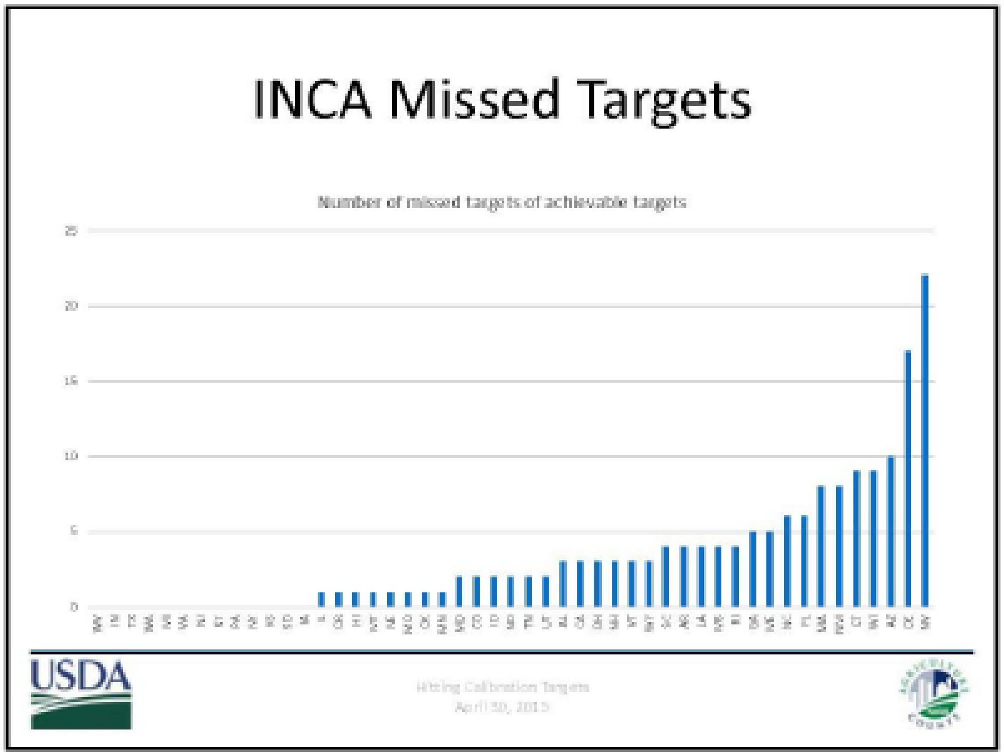


States

States with all possible targets attained	States with 1 - 5 missed targets			States with 5 - 10 missed targets	States with > 10 missed targets
IN, NY	MN	ID	RI	MA	NV
KY, WA	IL	HI	ME	FL	DE
IA	OR	MT	UT	NM	
KS	SC	LA		WI	
SD	MD	NE		NC	
WV	AR	OH		CT	
VA	AL	MO		AZ	
KS	GA	OK			
PA	CA	NH			
NJ	CO	ND			
TX	MS	WY			
MI	TN	VT			

USDA

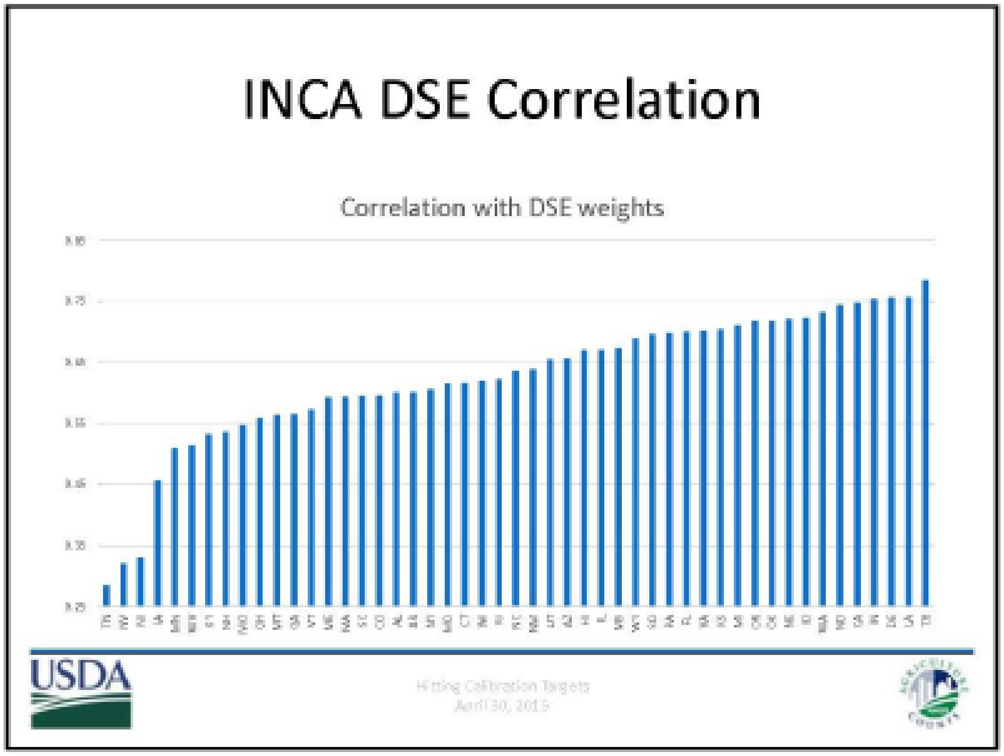
Hitting Calibration Targets
April 30, 2015



INCA Missed Targets

Number of missed targets of achievable targets

Hitting Calibration Targets
April 30, 2015

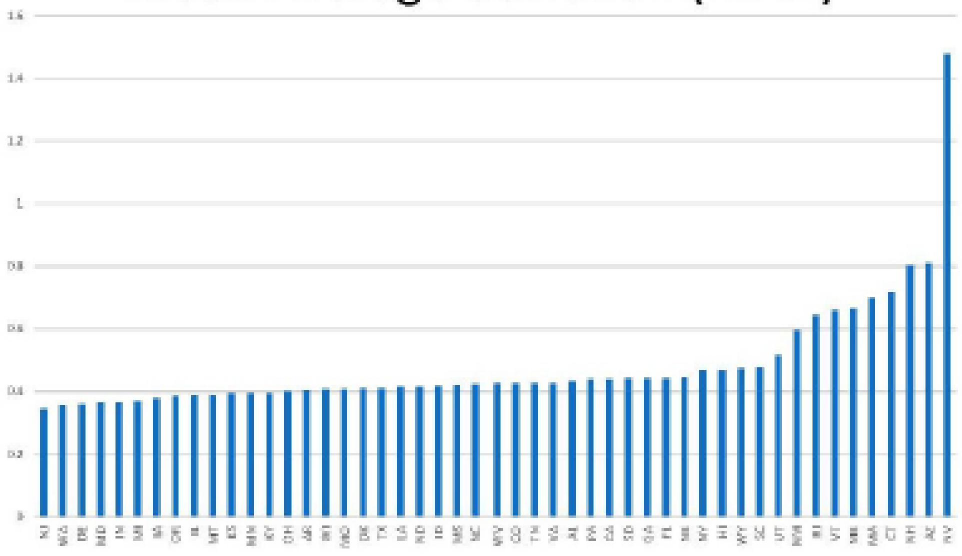


INCA DSE Correlation

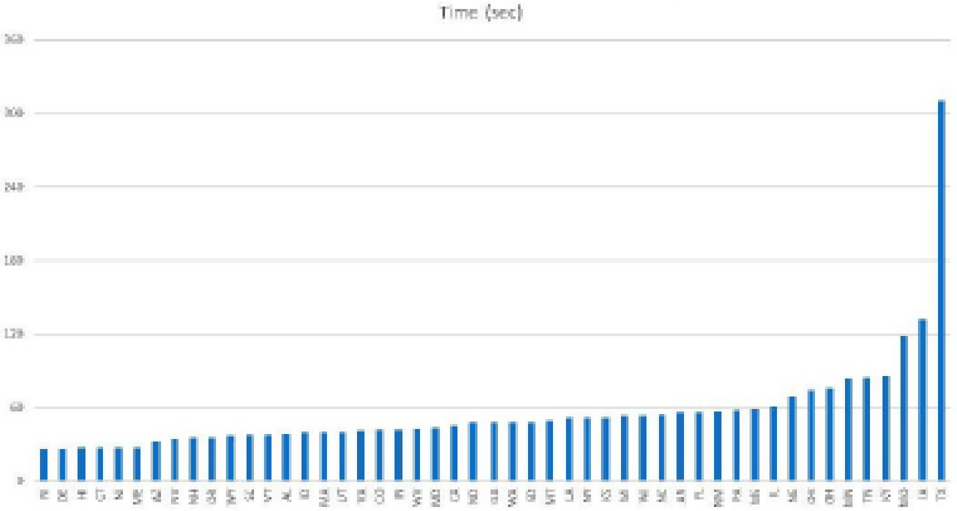
Correlation with DSE weights

Hitting Calibration Targets
April 30, 2015

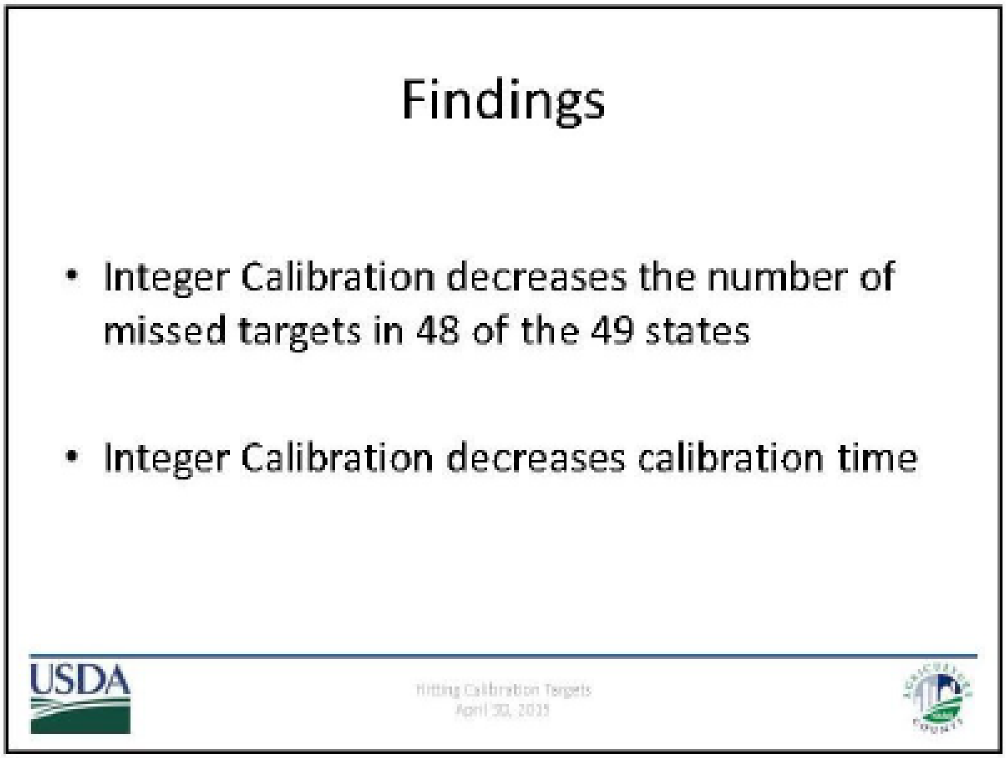
INCA
Mean Average deviation (MAD)



INCA Computational Speed

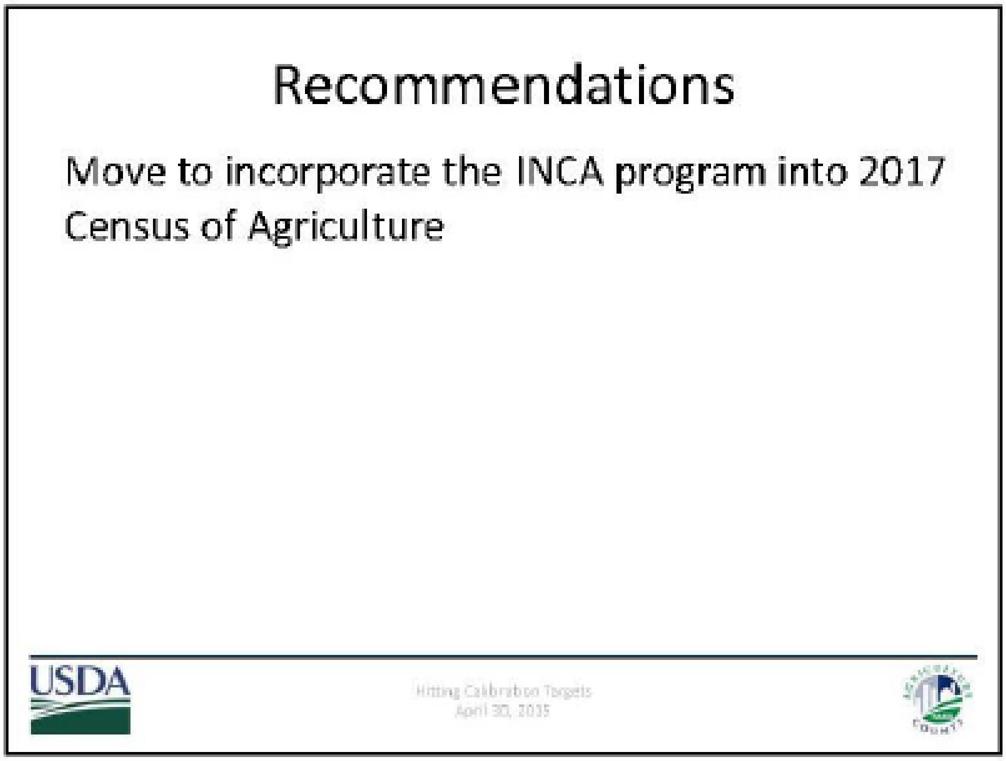


- Average time per state using old code is 30 mins



Findings

- Integer Calibration decreases the number of missed targets in 48 of the 49 states
- Integer Calibration decreases calibration time



Recommendations

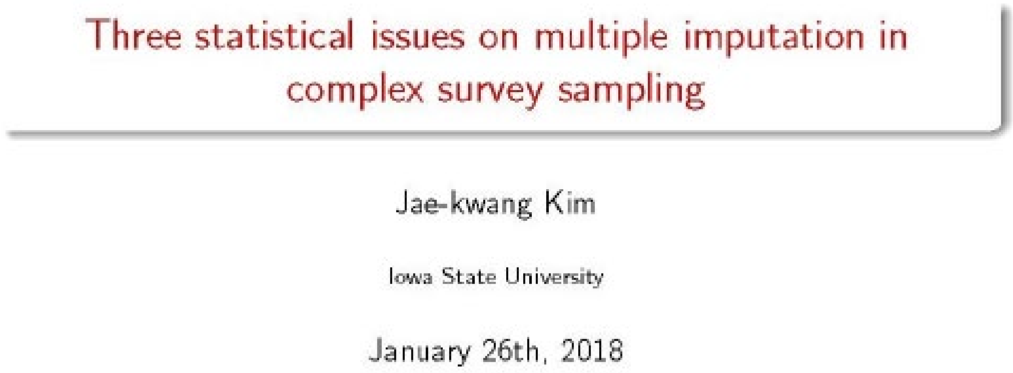
Move to incorporate the INCA program into 2017 Census of Agriculture

Thank you!

Luca Sartore, PhD - Luca.Sartore@nass.usda.gov
Kelly Toppin, PhD - Kelly.Toppin@nass.usda.gov
Clifford Spiegelman, PhD - Cliff@stat.tamu.edu

Hitting Calibration Targets
April 30, 2015

Three Statistical Issues on Multiple Imputation in Complex Survey Sampling

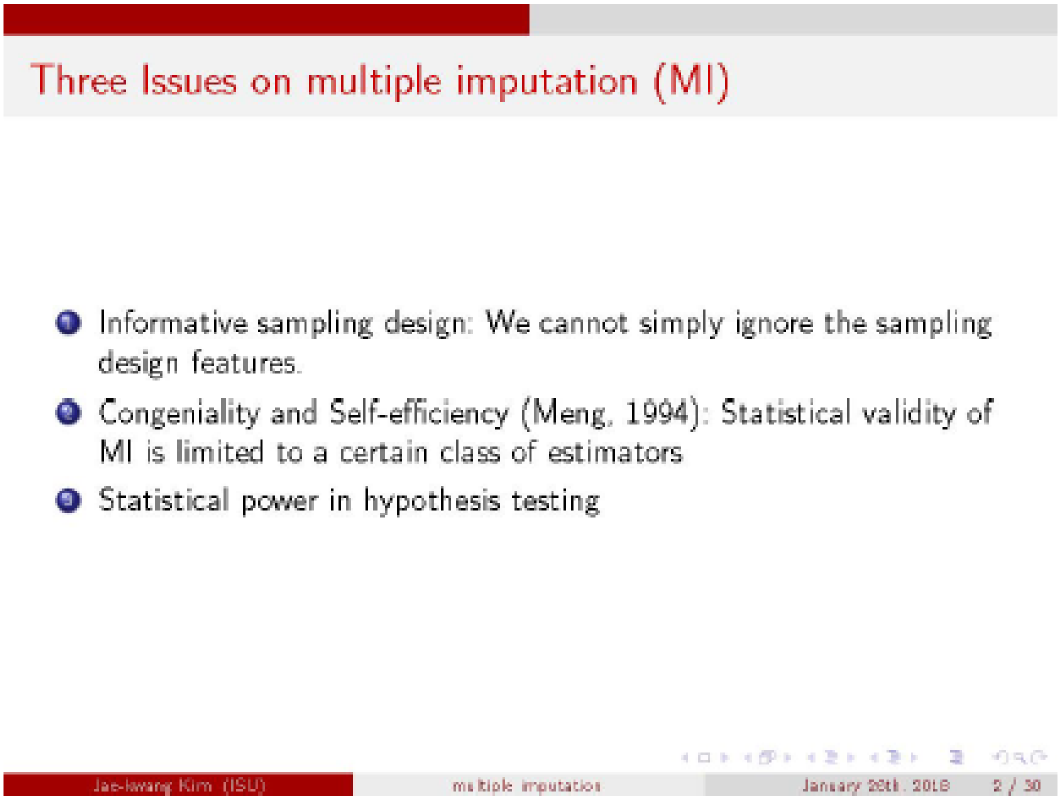


Three statistical issues on multiple imputation in complex survey sampling

Jae-kwang Kim

Iowa State University

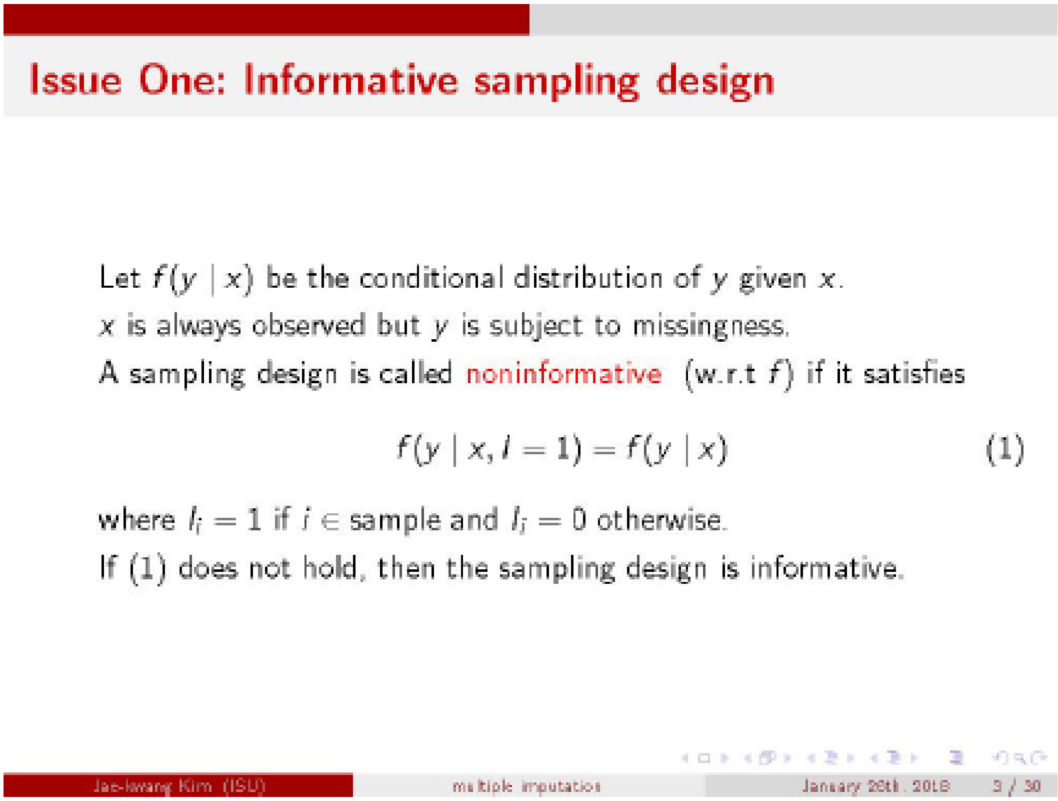
January 26th, 2018



Three Issues on multiple imputation (MI)

1. Informative sampling design: We cannot simply ignore the sampling design features.
2. Congeniality and Self-efficiency (Meng, 1994): Statistical validity of MI is limited to a certain class of estimators
3. Statistical power in hypothesis testing

Jae-kwang Kim (ISU) multiple imputation January 26th, 2018 2 / 30



Issue One: Informative sampling design

Let $f(y \mid x)$ be the conditional distribution of y given x.

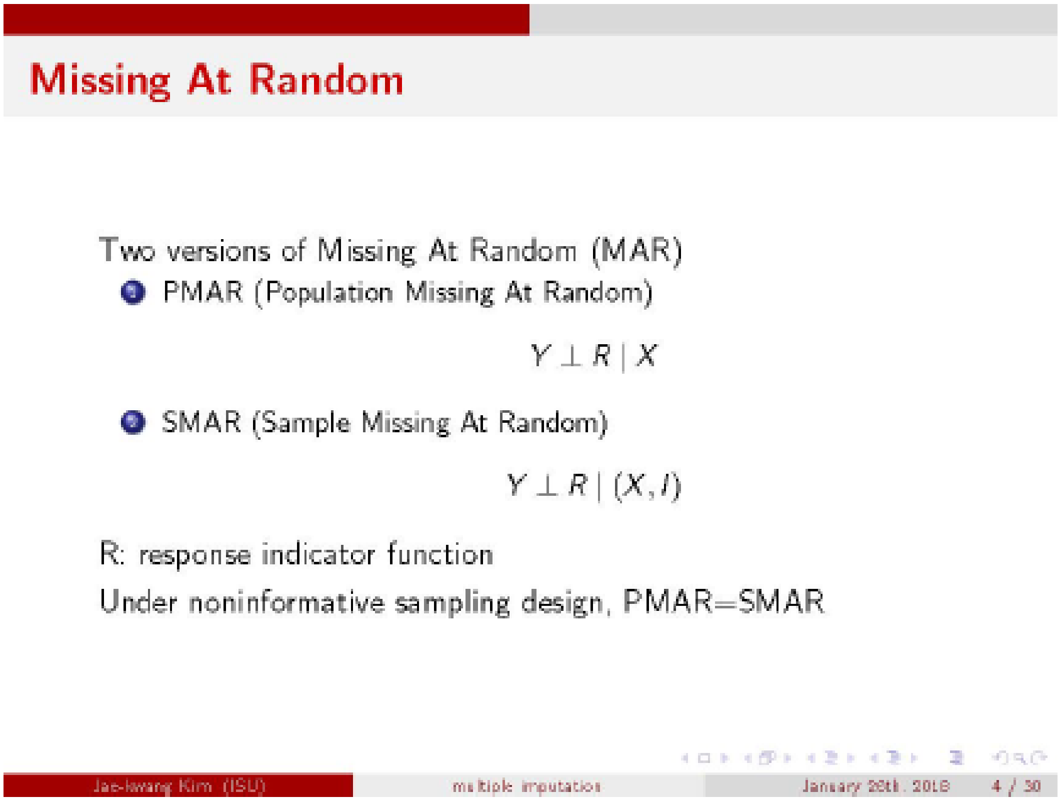
x is always observed but y is subject to missingness.

A sampling design is called noninformative (w.r.t f) if it satisfies

$$f(y \mid x, I = 1) = f(y \mid x) \qquad (1)$$

where $I_i = 1$ if $i \in$ sample and $I_i = 0$ otherwise.

If (1) does not hold, then the sampling design is informative.



Missing At Random

Two versions of Missing At Random (MAR)

1. PMAR (Population Missing At Random)

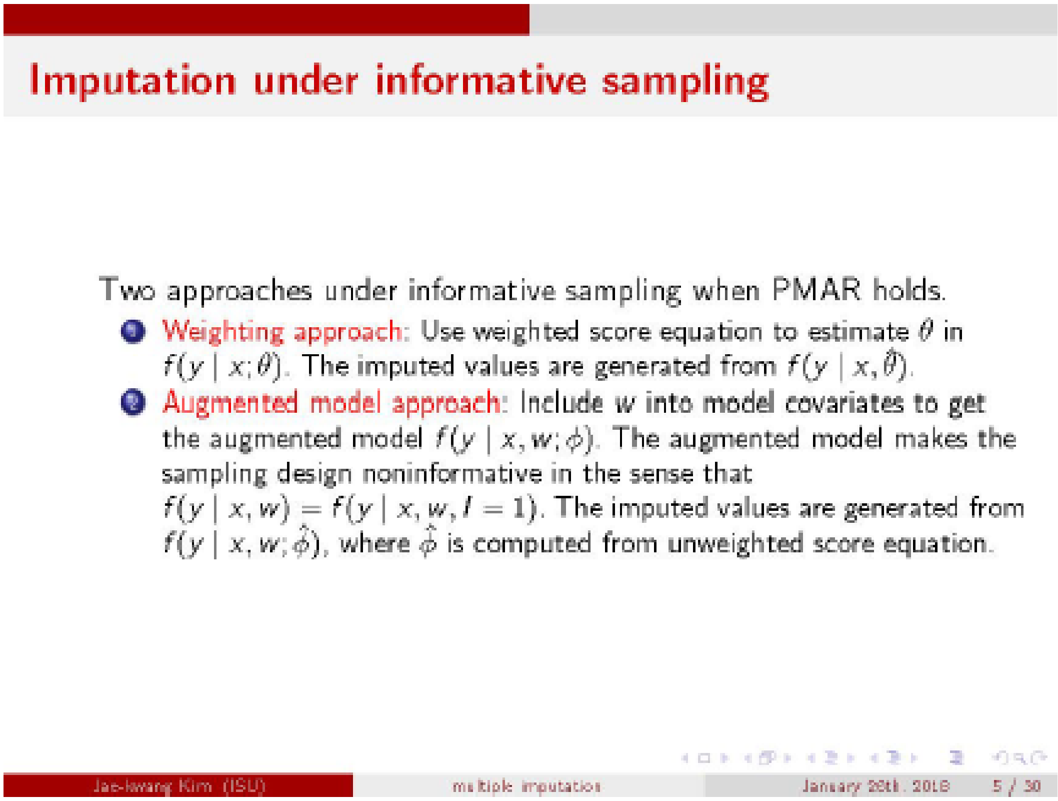
$$Y \perp R \mid X$$

2. SMAR (Sample Missing At Random)

$$Y \perp R \mid (X, I)$$

R: response indicator function

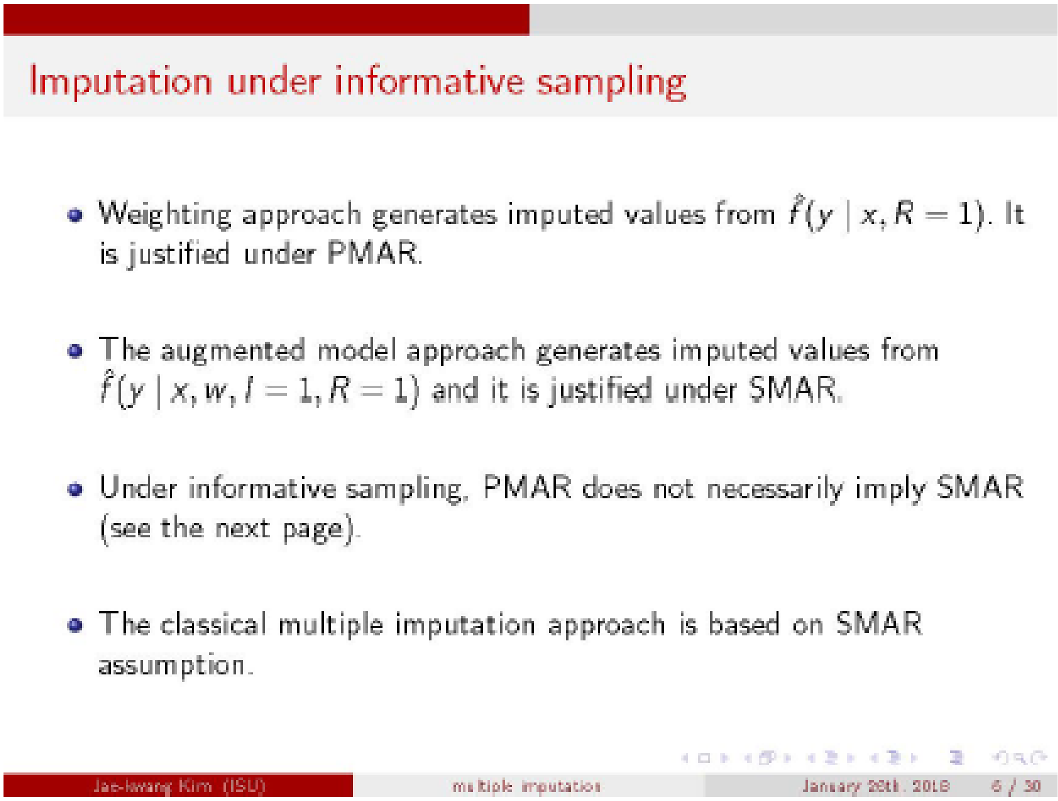
Under noninformative sampling design, PMAR=SMAR



Imputation under informative sampling

Two approaches under informative sampling when PMAR holds.

1. Weighting approach: Use weighted score equation to estimate θ in $f(y \mid x; \theta)$. The imputed values are generated from $f(y \mid x, \hat{\theta})$.
2. Augmented model approach: Include w into model covariates to get the augmented model $f(y \mid x, w; \phi)$. The augmented model makes the sampling design noninformative in the sense that $f(y \mid x, w) = f(y \mid x, w, I = 1)$. The imputed values are generated from $f(y \mid x, w; \hat{\phi})$, where $\hat{\phi}$ is computed from unweighted score equation.

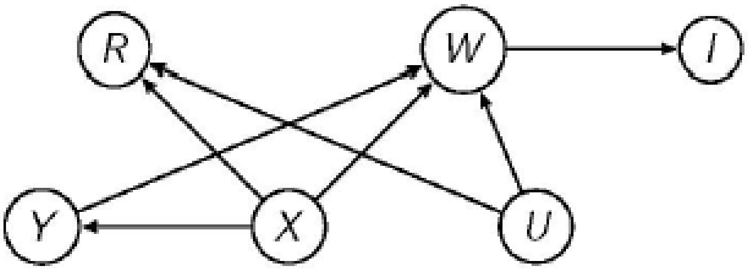


Imputation under informative sampling

- Weighting approach generates imputed values from $\hat{f}(y \mid x, R = 1)$. It is justified under PMAR.
- The augmented model approach generates imputed values from $\hat{f}(y \mid x, w, I = 1, R = 1)$ and it is justified under SMAR.
- Under informative sampling, PMAR does not necessarily imply SMAR (see the next page).
- The classical multiple imputation approach is based on SMAR assumption.

Berg, Kim, and Skinner (2016; JSSAM)

Figure: A Directed Acyclic Graph (DAG) for a setup where PMAR holds but SMAR does not hold. Variable U is latent in the sense that it is never observed.



$f(y \mid x, R) = f(y \mid x)$ holds but $f(y \mid x, w, R) \neq f(y \mid x, w)$.

MI under informative sampling

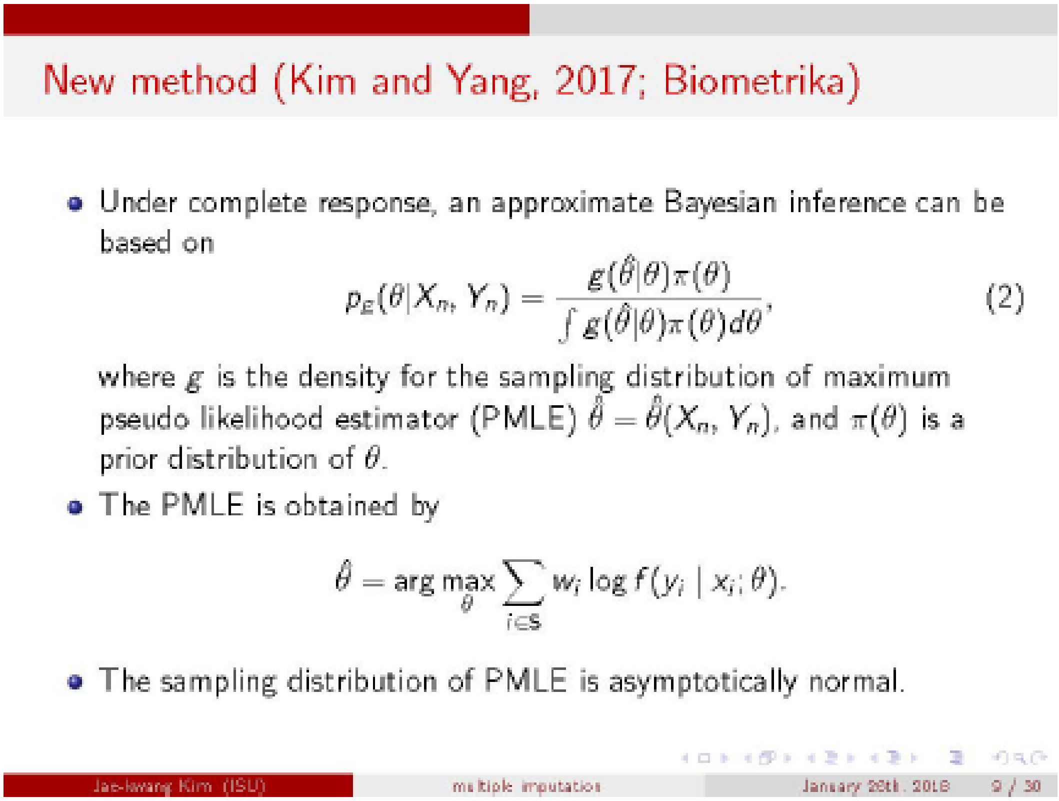
- Under informative sampling, the sample distribution is different from the population distribution which follows from the marginal sample distribution,

$$f(y_i|x_i, I_i = 1) = \frac{P(I_i = 1|x_i, y_i)f(y_i|x_i)}{P(I_i = 1|x_i)}.$$

- Recall that the posterior distribution for multiple imputation is

$$p(\theta|X_n, Y_{\text{obs}}) = \frac{\int L_s(\theta|X_n, Y_n)\pi(\theta)dY_{\text{mis}}}{\int\int L_s(\theta|X_n, Y_n)\pi(\theta)dY_{\text{mis}}d\theta}.$$

- So, it is difficult to obtain the likelihood function $L_s(\theta|X_n, Y_n)$ directly from the population distribution.



New method (Kim and Yang, 2017; Biometrika)

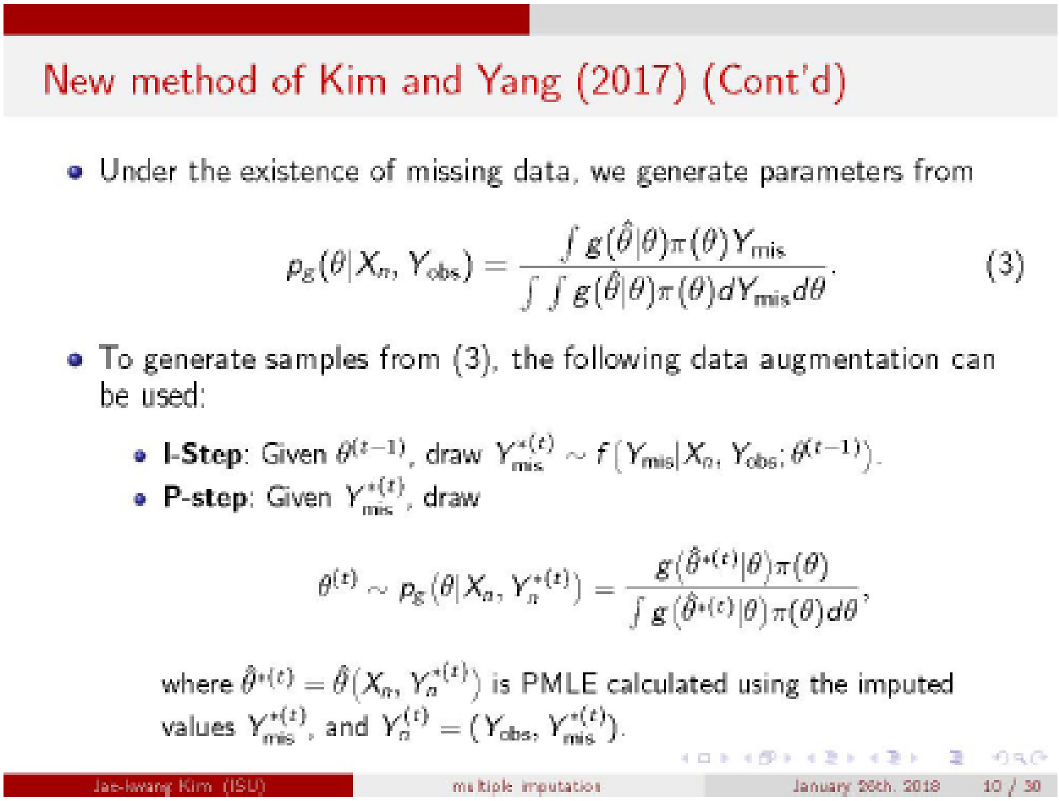
- Under complete response, an approximate Bayesian inference can be based on

$$p_g(\theta|X_n, Y_n) = \frac{g(\hat{\theta}|\theta)\pi(\theta)}{\int g(\hat{\theta}|\theta)\pi(\theta)d\theta}, \qquad (2)$$

  where g is the density for the sampling distribution of maximum pseudo likelihood estimator (PMLE) $\hat{\theta} = \hat{\theta}(X_n, Y_n)$, and $\pi(\theta)$ is a prior distribution of θ.
- The PMLE is obtained by

$$\hat{\theta} = \arg\max_{\theta} \sum_{i \in S} w_i \log f(y_i \mid x_i; \theta).$$

- The sampling distribution of PMLE is asymptotically normal.



New method of Kim and Yang (2017) (Cont'd)

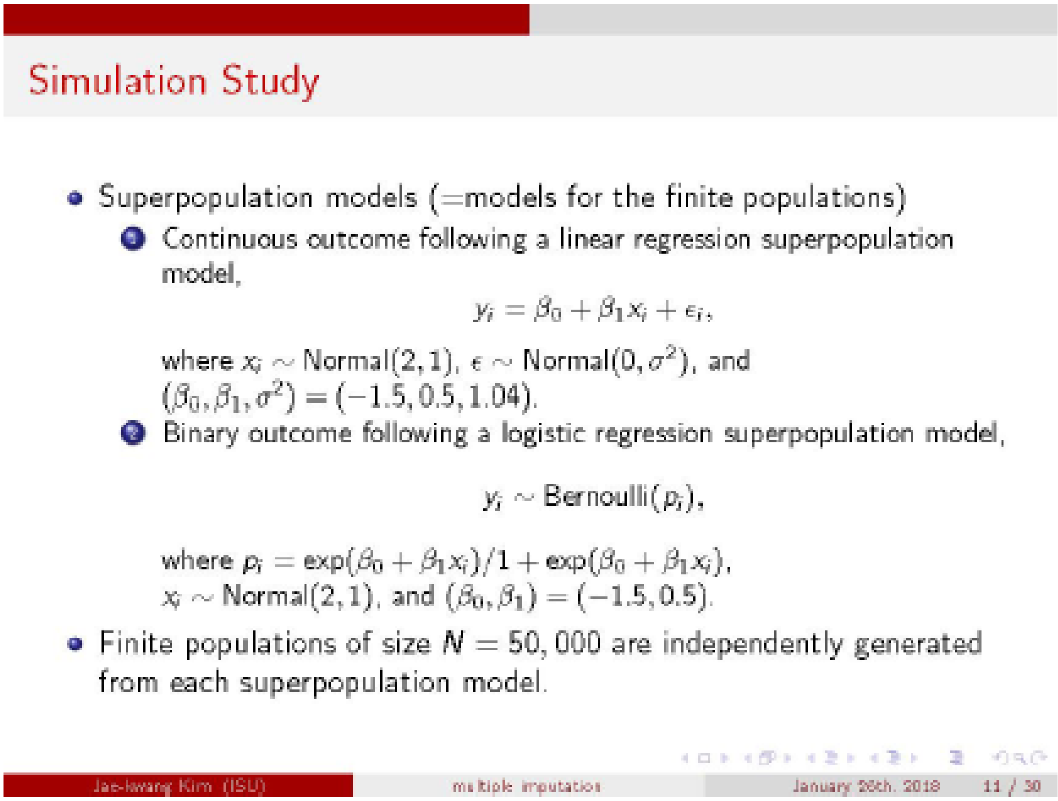
- Under the existence of missing data, we generate parameters from

$$p_g(\theta|X_n, Y_{\rm obs}) = \frac{\int g(\hat{\theta}|\theta)\pi(\theta)Y_{\rm mis}}{\int\int g(\hat{\theta}|\theta)\pi(\theta)dY_{\rm mis}d\theta}. \qquad (3)$$

- To generate samples from (3), the following data augmentation can be used:
  - **I-Step**: Given $\theta^{(t-1)}$, draw $Y_{\rm mis}^{*(t)} \sim f\left(Y_{\rm mis}|X_n, Y_{\rm obs}; \theta^{(t-1)}\right)$.
  - **P-step**: Given $Y_{\rm mis}^{*(t)}$, draw

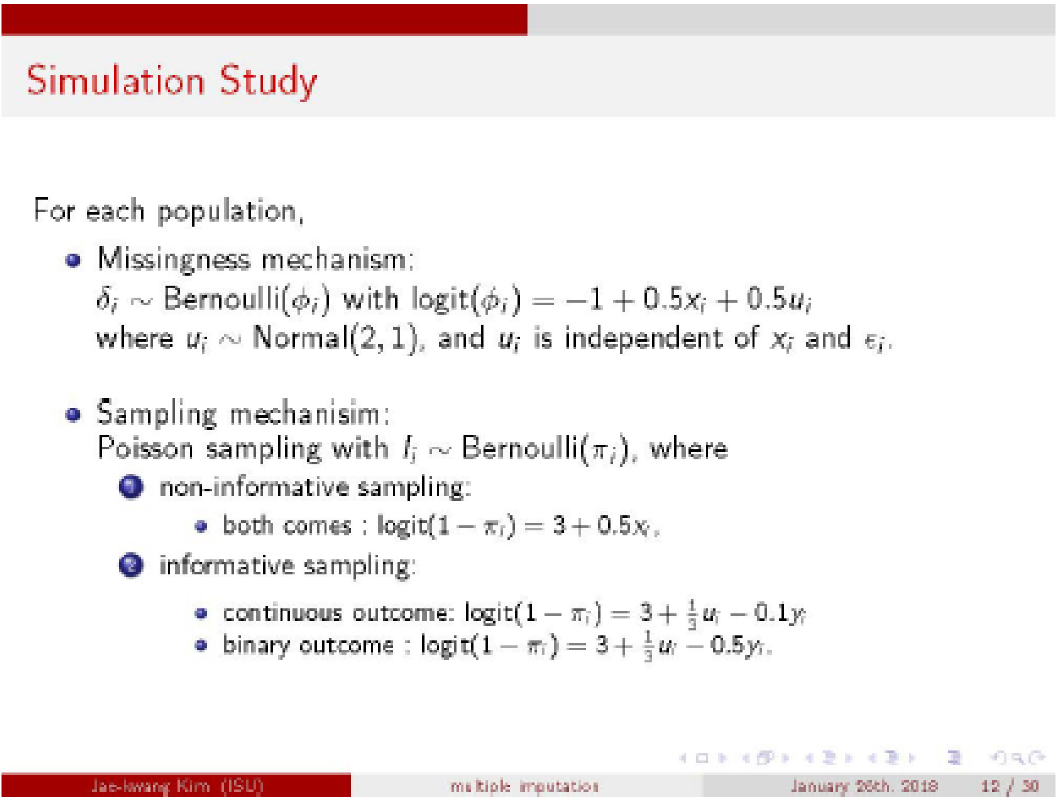
$$\theta^{(t)} \sim p_g\left(\theta|X_n, Y_n^{*(t)}\right) = \frac{g\left(\hat{\theta}^{*(t)}|\theta\right)\pi(\theta)}{\int g\left(\hat{\theta}^{*(t)}|\theta\right)\pi(\theta)d\theta},$$

    where $\hat{\theta}^{*(t)} = \hat{\theta}\left(X_n, Y_n^{*(t)}\right)$ is PMLE calculated using the imputed values $Y_{\rm mis}^{*(t)}$, and $Y_n^{(t)} = \left(Y_{\rm obs}, Y_{\rm mis}^{*(t)}\right)$.



Simulation Study

- Superpopulation models (=models for the finite populations)
  1. Continuous outcome following a linear regression superpopulation model,
     $$y_i = \beta_0 + \beta_1 x_i + \epsilon_i,$$
     where $x_i \sim \text{Normal}(2,1)$, $\epsilon \sim \text{Normal}(0,\sigma^2)$, and $(\beta_0, \beta_1, \sigma^2) = (-1.5, 0.5, 1.04)$.
  2. Binary outcome following a logistic regression superpopulation model,
     $$y_i \sim \text{Bernoulli}(p_i),$$
     where $p_i = \exp(\beta_0 + \beta_1 x_i)/1 + \exp(\beta_0 + \beta_1 x_i)$, $x_i \sim \text{Normal}(2,1)$, and $(\beta_0, \beta_1) = (-1.5, 0.5)$.
- Finite populations of size $N = 50,000$ are independently generated from each superpopulation model.

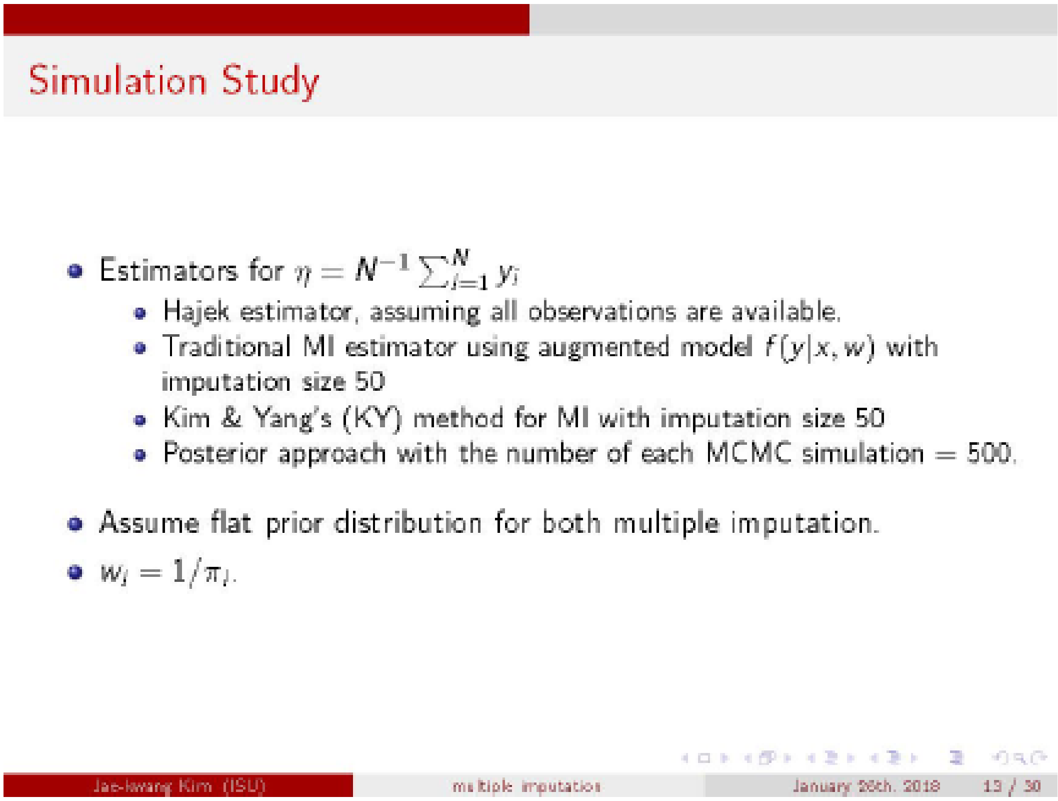


Simulation Study

For each population,

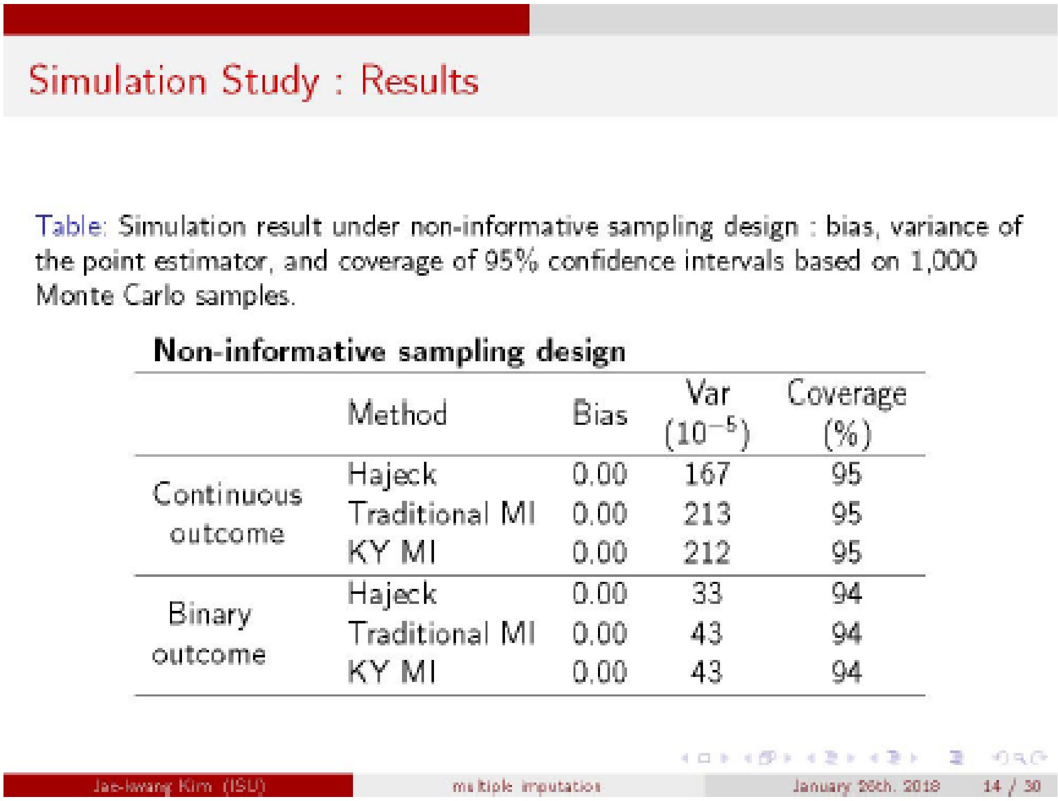
- Missingness mechanism:
  $\delta_i \sim \text{Bernoulli}(\phi_i)$ with $\text{logit}(\phi_i) = -1 + 0.5x_i + 0.5u_i$
  where $u_i \sim \text{Normal}(2,1)$, and u_i is independent of x_i and ϵ_i.
- Sampling mechanisim:
  Poisson sampling with $I_i \sim \text{Bernoulli}(\pi_i)$, where
  1. non-informative sampling:
     - both comes : $\text{logit}(1-\pi_i) = 3 + 0.5x_i$.
  2. informative sampling:
     - continuous outcome: $\text{logit}(1-\pi_i) = 3 + \frac{1}{3}u_i - 0.1y_i$
     - binary outcome : $\text{logit}(1-\pi_i) = 3 + \frac{1}{3}u_i - 0.5y_i$.

5



Simulation Study

- Estimators for $\eta = N^{-1}\sum_{i=1}^{N} y_i$
  - Hajek estimator, assuming all observations are available.
  - Traditional MI estimator using augmented model $f(y|x,w)$ with imputation size 50
  - Kim & Yang's (KY) method for MI with imputation size 50
  - Posterior approach with the number of each MCMC simulation = 500.
- Assume flat prior distribution for both multiple imputation.
- $w_i = 1/\pi_i$.

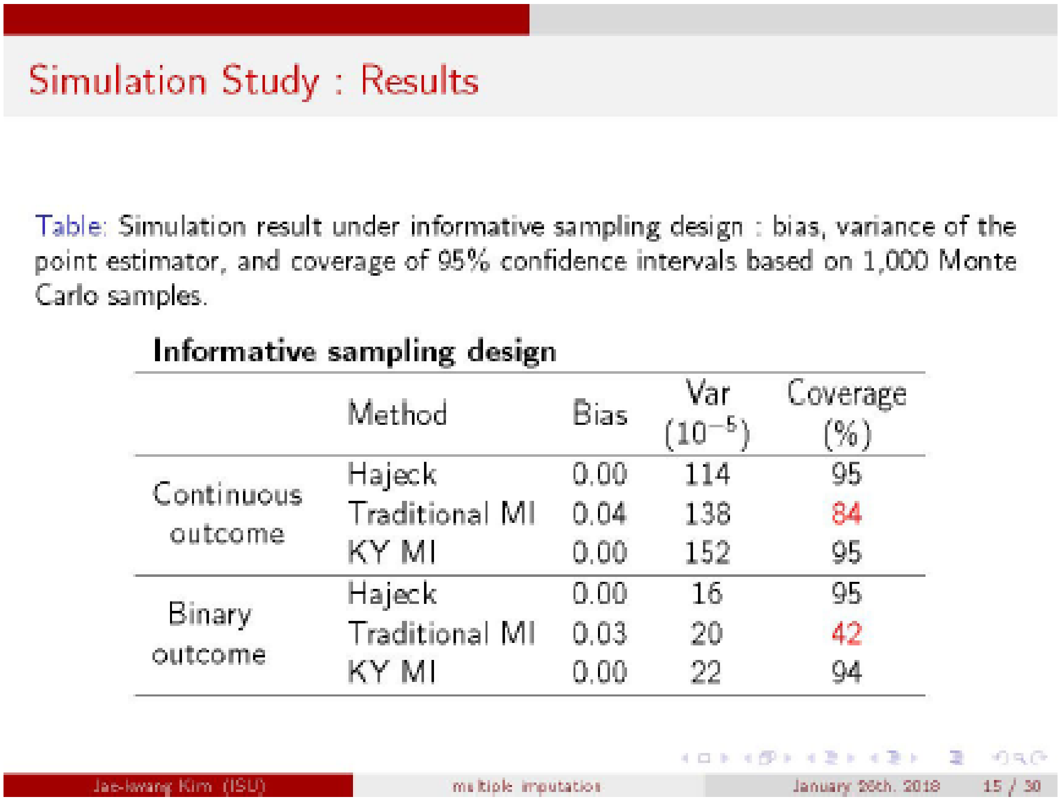


Simulation Study : Results

Table: Simulation result under non-informative sampling design : bias, variance of the point estimator, and coverage of 95% confidence intervals based on 1,000 Monte Carlo samples.

Non-informative sampling design

	Method	Bias	Var (10^{-5})	Coverage (%)
Continuous outcome	Hajeck	0.00	167	95
	Traditional MI	0.00	213	95
	KY MI	0.00	212	95
Binary outcome	Hajeck	0.00	33	94
	Traditional MI	0.00	43	94
	KY MI	0.00	43	94

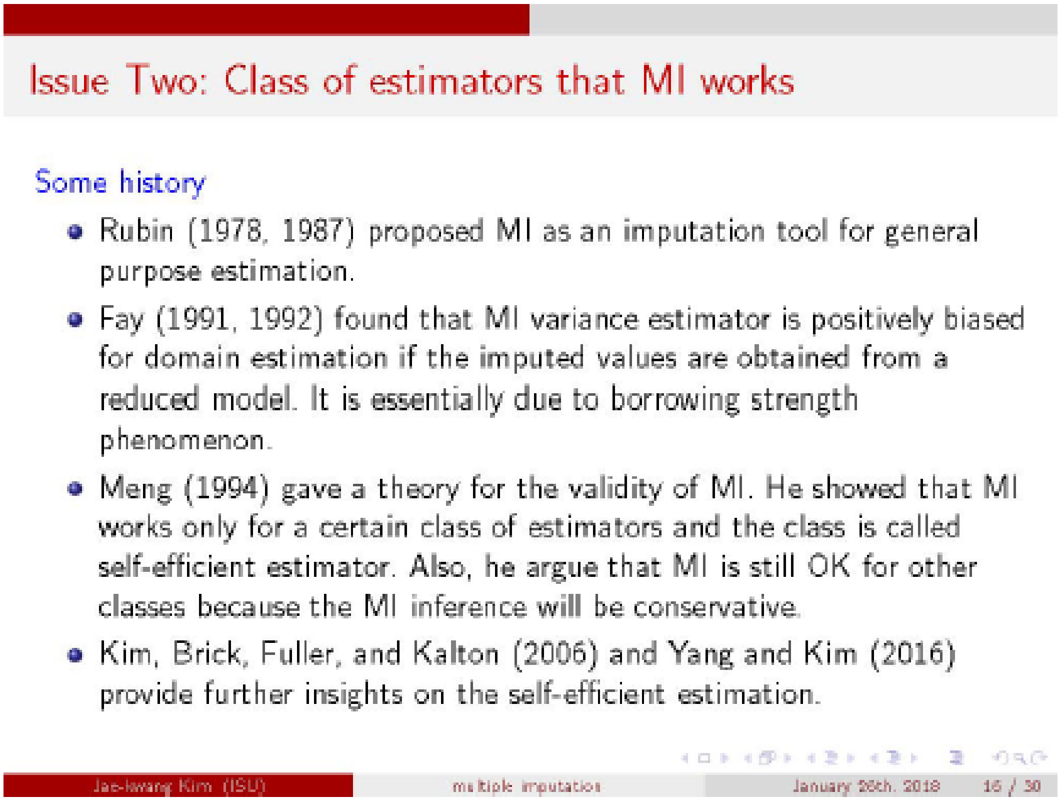


Simulation Study : Results

Table: Simulation result under informative sampling design : bias, variance of the point estimator, and coverage of 95% confidence intervals based on 1,000 Monte Carlo samples.

Informative sampling design

	Method	Bias	Var (10^{-5})	Coverage (%)
Continuous outcome	Hajeck	0.00	114	95
	Traditional MI	0.04	138	84
	KY MI	0.00	152	95
Binary outcome	Hajeck	0.00	16	95
	Traditional MI	0.03	20	42
	KY MI	0.00	22	94



Issue Two: Class of estimators that MI works

Some history

- Rubin (1978, 1987) proposed MI as an imputation tool for general purpose estimation.
- Fay (1991, 1992) found that MI variance estimator is positively biased for domain estimation if the imputed values are obtained from a reduced model. It is essentially due to borrowing strength phenomenon.
- Meng (1994) gave a theory for the validity of MI. He showed that MI works only for a certain class of estimators and the class is called self-efficient estimator. Also, he argue that MI is still OK for other classes because the MI inference will be conservative.
- Kim, Brick, Fuller, and Kalton (2006) and Yang and Kim (2016) provide further insights on the self-efficient estimation.

Numerical illustration

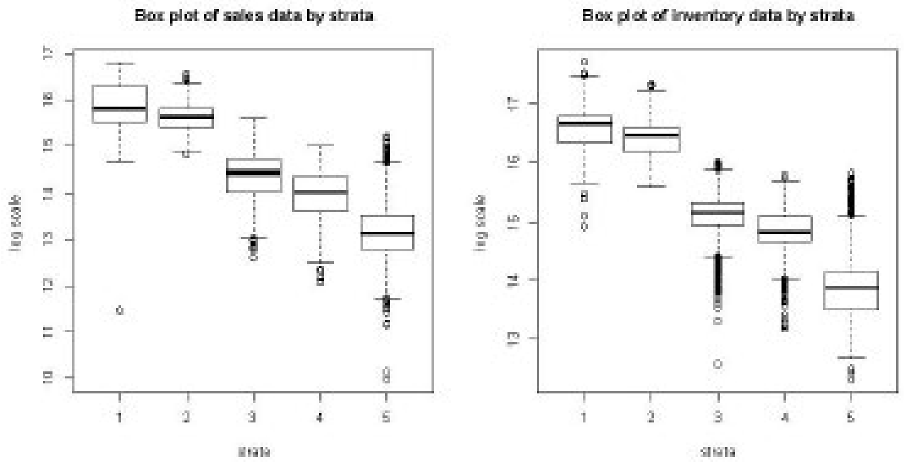
A pseudo finite population constructed from a single month data in Monthly Retail Trade Survey (MRTS) at US Bureau of Census

$N = 7,260$ retail business units in five strata

Three variables in the data

- h: stratum
- x_{hi}: inventory values
- y_{hi}: sales

Box plot of log sales and log inventory values by strata



Imputation model

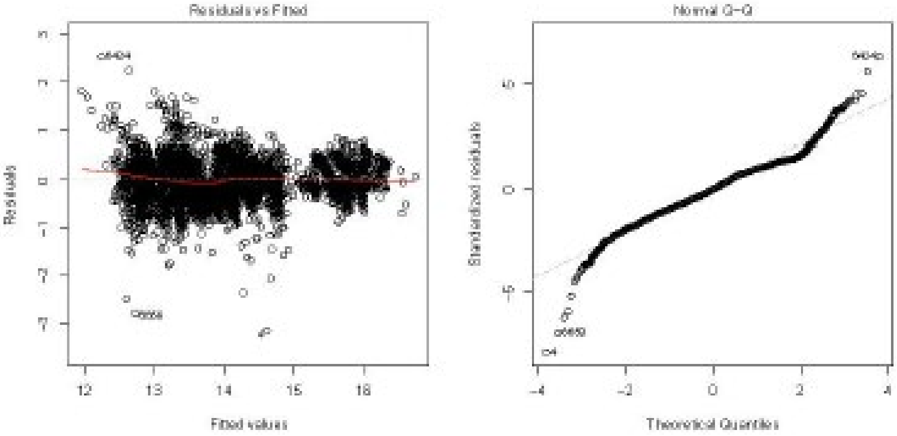
$$log(y_{hi}) = \beta_{0h} + \beta_1 \log(x_{hi}) + e_{hi}$$

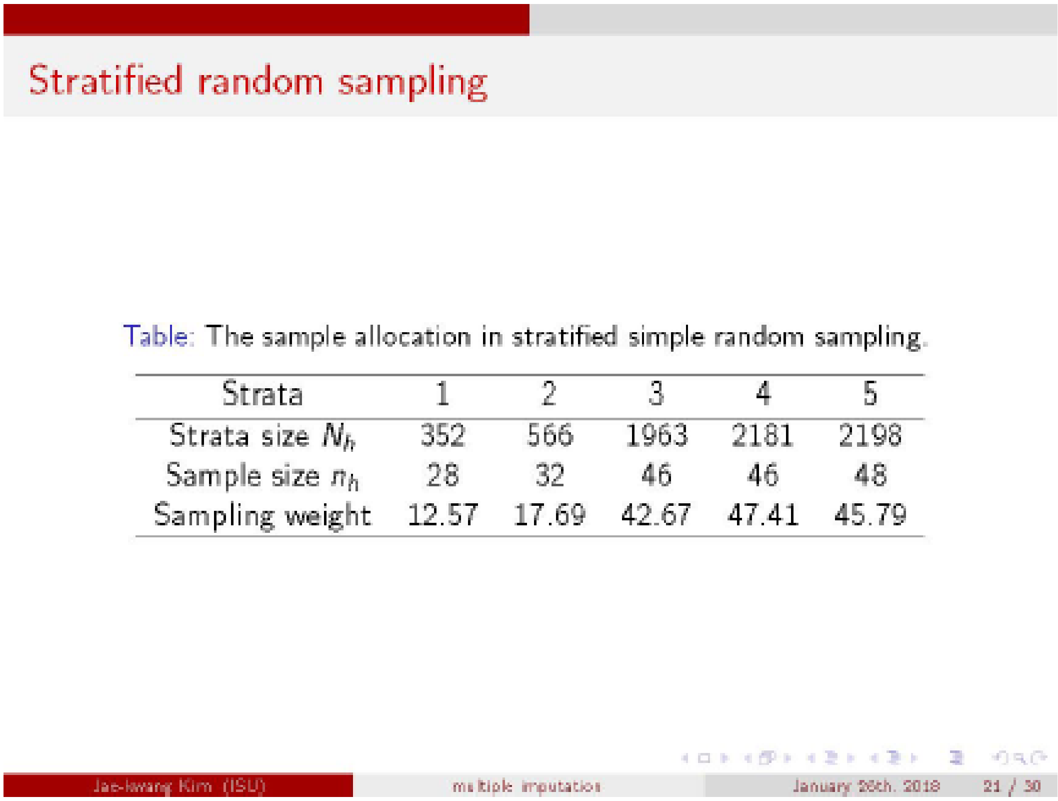
where

$$e_{hi} \sim N(0, \sigma^2)$$

Residual plot and residual QQ plot

Regression model of log(y) against log(x) and strata indicator

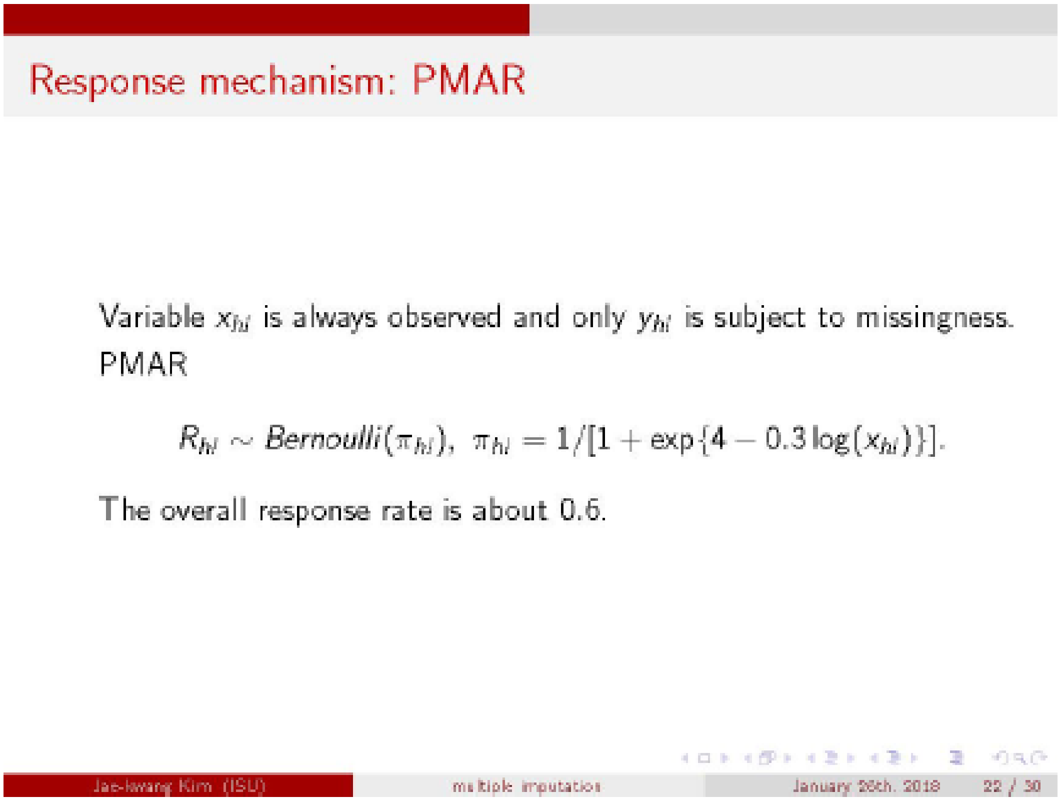




Stratified random sampling

Table: The sample allocation in stratified simple random sampling.

Strata	1	2	3	4	5
Strata size N_h	352	566	1963	2181	2198
Sample size n_h	28	32	46	46	48
Sampling weight	12.57	17.69	42.67	47.41	45.79

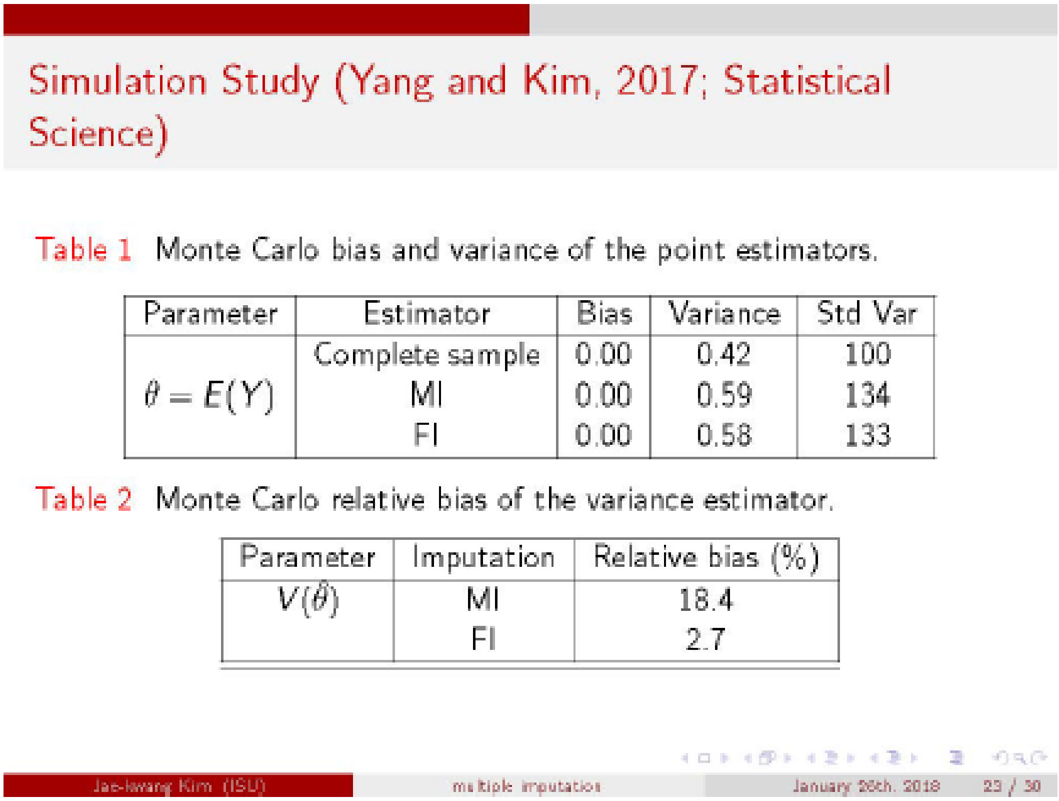


Response mechanism: PMAR

Variable x_{hi} is always observed and only y_{hi} is subject to missingness.
PMAR

$$R_{hi} \sim Bernoulli(\pi_{hi}), \ \ \pi_{hi} = 1/[1 + \exp\{4 - 0.3 \log(x_{hi})\}].$$

The overall response rate is about 0.6.



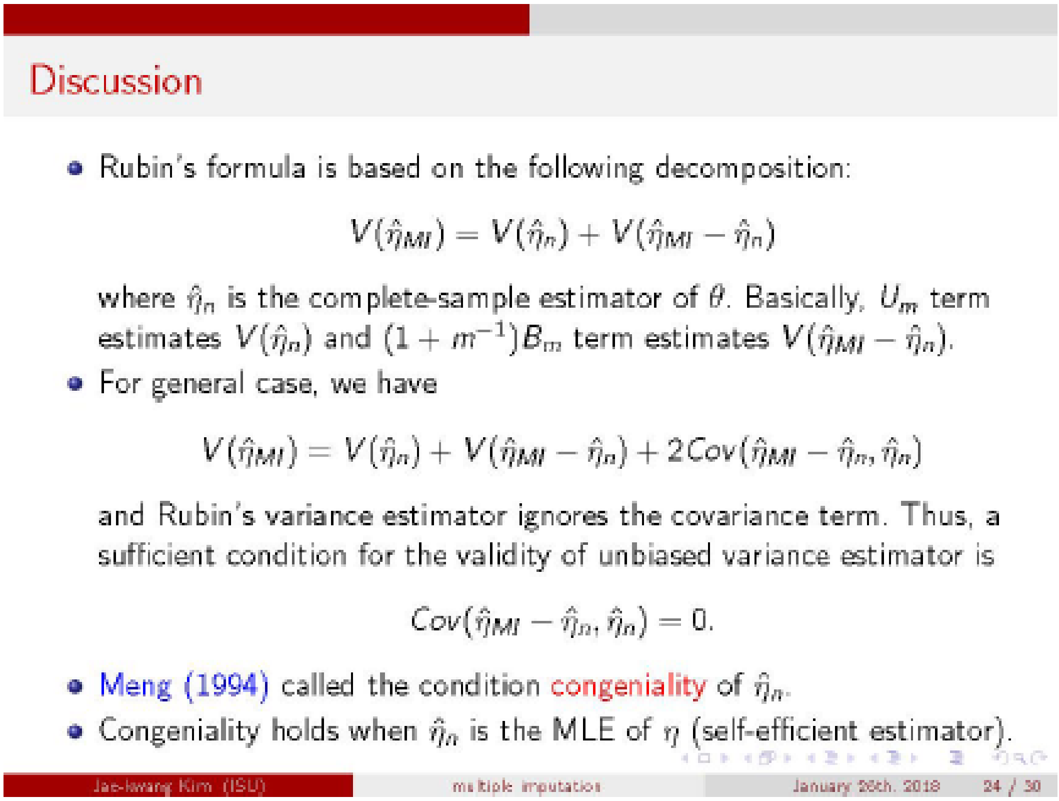
Simulation Study (Yang and Kim, 2017; Statistical Science)

Table 1 Monte Carlo bias and variance of the point estimators.

Parameter	Estimator	Bias	Variance	Std Var
$\theta = E(Y)$	Complete sample	0.00	0.42	100
	MI	0.00	0.59	134
	FI	0.00	0.58	133

Table 2 Monte Carlo relative bias of the variance estimator.

Parameter	Imputation	Relative bias (%)
$V(\hat{\theta})$	MI	18.4
	FI	2.7



Discussion

- Rubin's formula is based on the following decomposition:

$$V(\hat{\eta}_{MI}) = V(\hat{\eta}_n) + V(\hat{\eta}_{MI} - \hat{\eta}_n)$$

where $\hat{\eta}_n$ is the complete-sample estimator of θ. Basically, U_m term estimates $V(\hat{\eta}_n)$ and $(1 + m^{-1})B_m$ term estimates $V(\hat{\eta}_{MI} - \hat{\eta}_n)$.

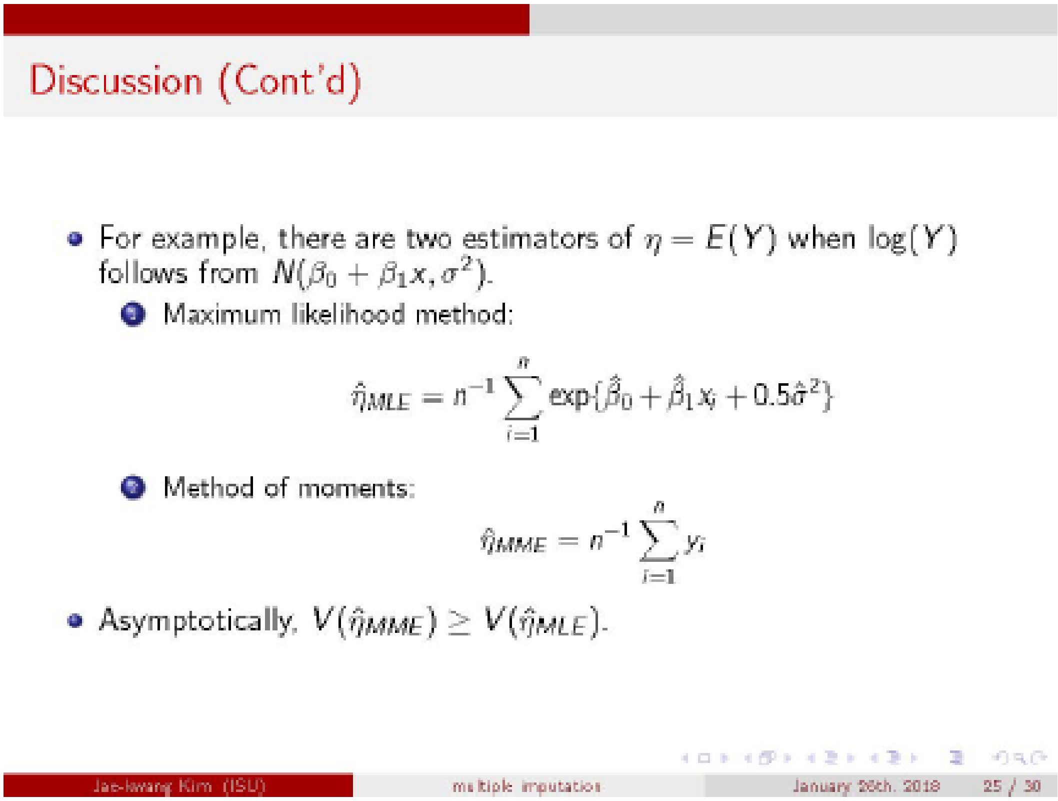
- For general case, we have

$$V(\hat{\eta}_{MI}) = V(\hat{\eta}_n) + V(\hat{\eta}_{MI} - \hat{\eta}_n) + 2Cov(\hat{\eta}_{MI} - \hat{\eta}_n, \hat{\eta}_n)$$

and Rubin's variance estimator ignores the covariance term. Thus, a sufficient condition for the validity of unbiased variance estimator is

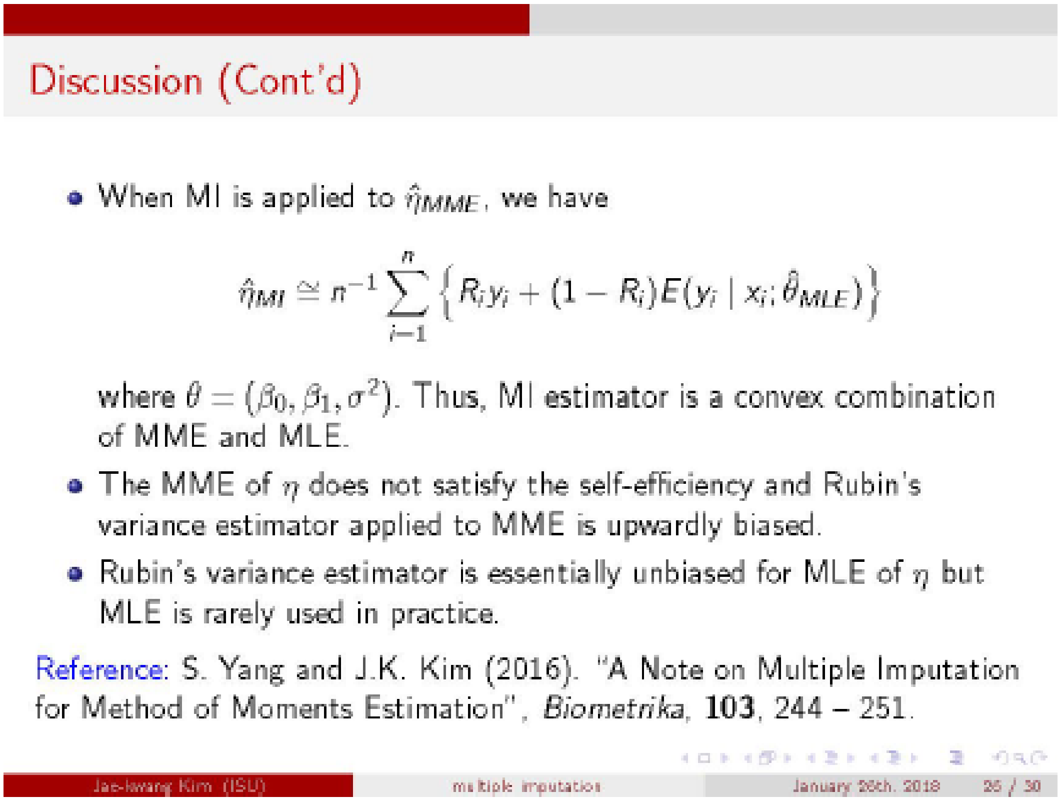
$$Cov(\hat{\eta}_{MI} - \hat{\eta}_n, \hat{\eta}_n) = 0.$$

- Meng (1994) called the condition congeniality of $\hat{\eta}_n$.
- Congeniality holds when $\hat{\eta}_n$ is the MLE of η (self-efficient estimator).



Discussion (Cont'd)

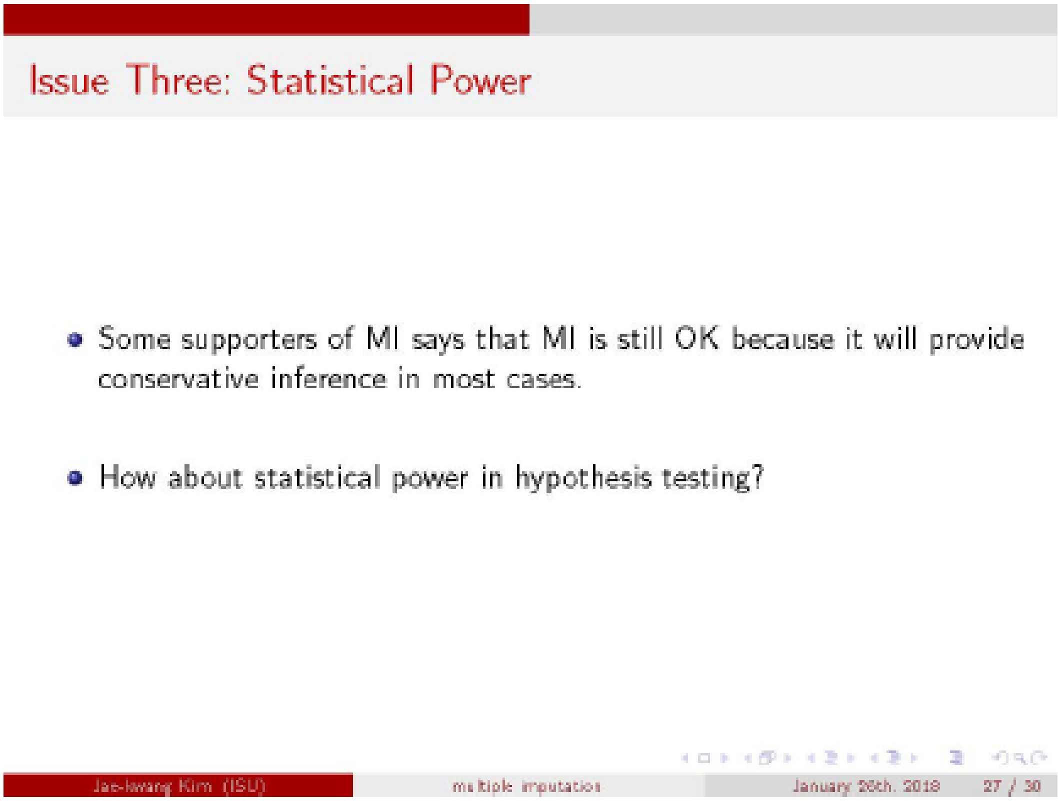
- For example, there are two estimators of $\eta = E(Y)$ when $\log(Y)$ follows from $N(\beta_0 + \beta_1 x, \sigma^2)$.
  1. Maximum likelihood method:
  $$\hat{\eta}_{MLE} = n^{-1}\sum_{i=1}^{n} \exp\{\hat{\beta}_0 + \hat{\beta}_1 x_i + 0.5\hat{\sigma}^2\}$$
  2. Method of moments:
  $$\hat{\eta}_{MME} = n^{-1}\sum_{i=1}^{n} y_i$$
- Asymptotically, $V(\hat{\eta}_{MME}) \geq V(\hat{\eta}_{MLE})$.



Discussion (Cont'd)

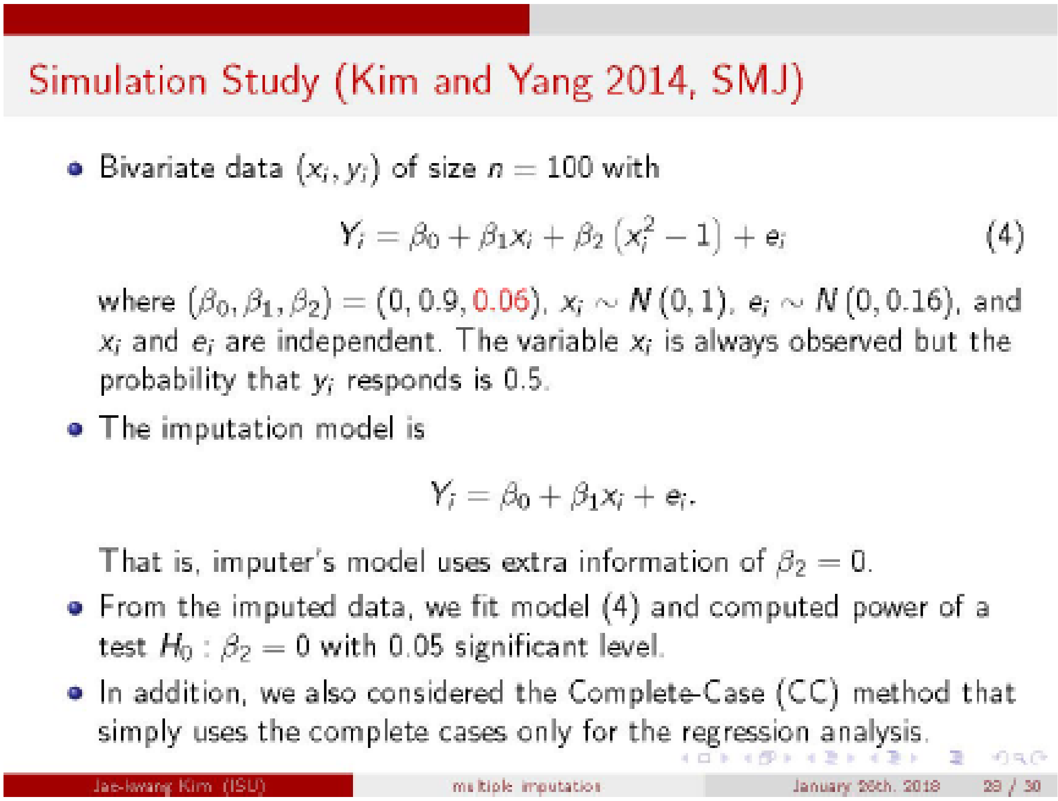
- When MI is applied to $\hat{\eta}_{MME}$, we have
$$\hat{\eta}_{MI} \cong n^{-1}\sum_{i=1}^{n}\left\{R_i y_i + (1-R_i)E(y_i \mid x_i; \hat{\theta}_{MLE})\right\}$$
where $\theta = (\beta_0, \beta_1, \sigma^2)$. Thus, MI estimator is a convex combination of MME and MLE.
- The MME of η does not satisfy the self-efficiency and Rubin's variance estimator applied to MME is upwardly biased.
- Rubin's variance estimator is essentially unbiased for MLE of η but MLE is rarely used in practice.

Reference: S. Yang and J.K. Kim (2016). "A Note on Multiple Imputation for Method of Moments Estimation", *Biometrika*, **103**, 244 – 251.



Issue Three: Statistical Power

- Some supporters of MI says that MI is still OK because it will provide conservative inference in most cases.
- How about statistical power in hypothesis testing?



Simulation Study (Kim and Yang 2014, SMJ)

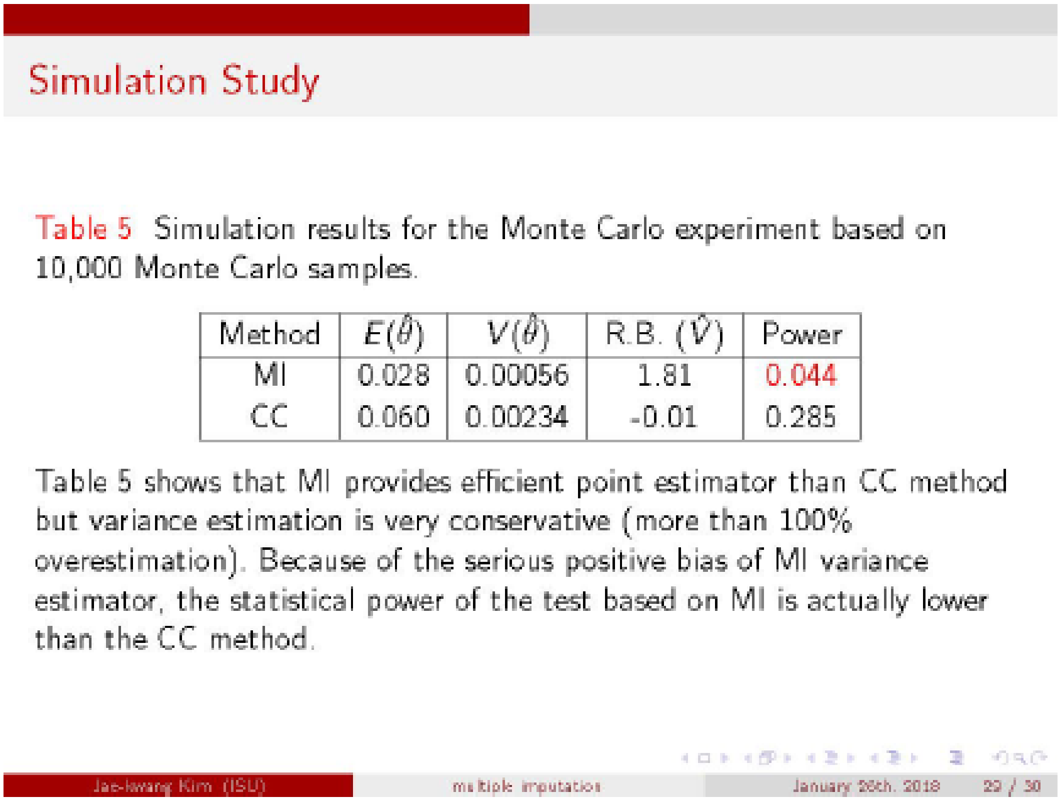
- Bivariate data (x_i, y_i) of size $n = 100$ with

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 \left(x_i^2 - 1\right) + e_i \tag{4}$$

  where $(\beta_0, \beta_1, \beta_2) = (0, 0.9, 0.06)$, $x_i \sim N(0, 1)$, $e_i \sim N(0, 0.16)$, and x_i and e_i are independent. The variable x_i is always observed but the probability that y_i responds is 0.5.
- The imputation model is

$$Y_i = \beta_0 + \beta_1 x_i + e_i.$$

  That is, imputer's model uses extra information of $\beta_2 = 0$.
- From the imputed data, we fit model (4) and computed power of a test $H_0 : \beta_2 = 0$ with 0.05 significant level.
- In addition, we also considered the Complete-Case (CC) method that simply uses the complete cases only for the regression analysis.

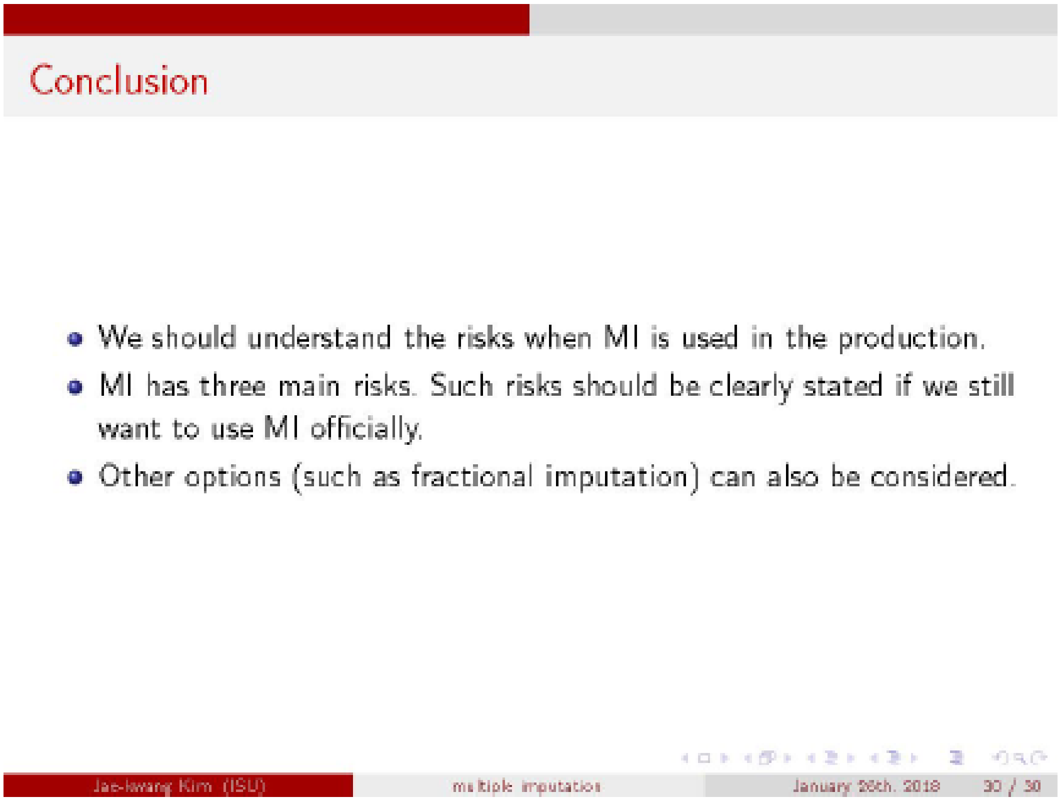


Simulation Study

Table 5 Simulation results for the Monte Carlo experiment based on 10,000 Monte Carlo samples.

Method	$E(\hat{\theta})$	$V(\hat{\theta})$	R.B. $(\hat{V})$	Power
MI	0.028	0.00056	1.81	0.044
CC	0.060	0.00234	-0.01	0.285

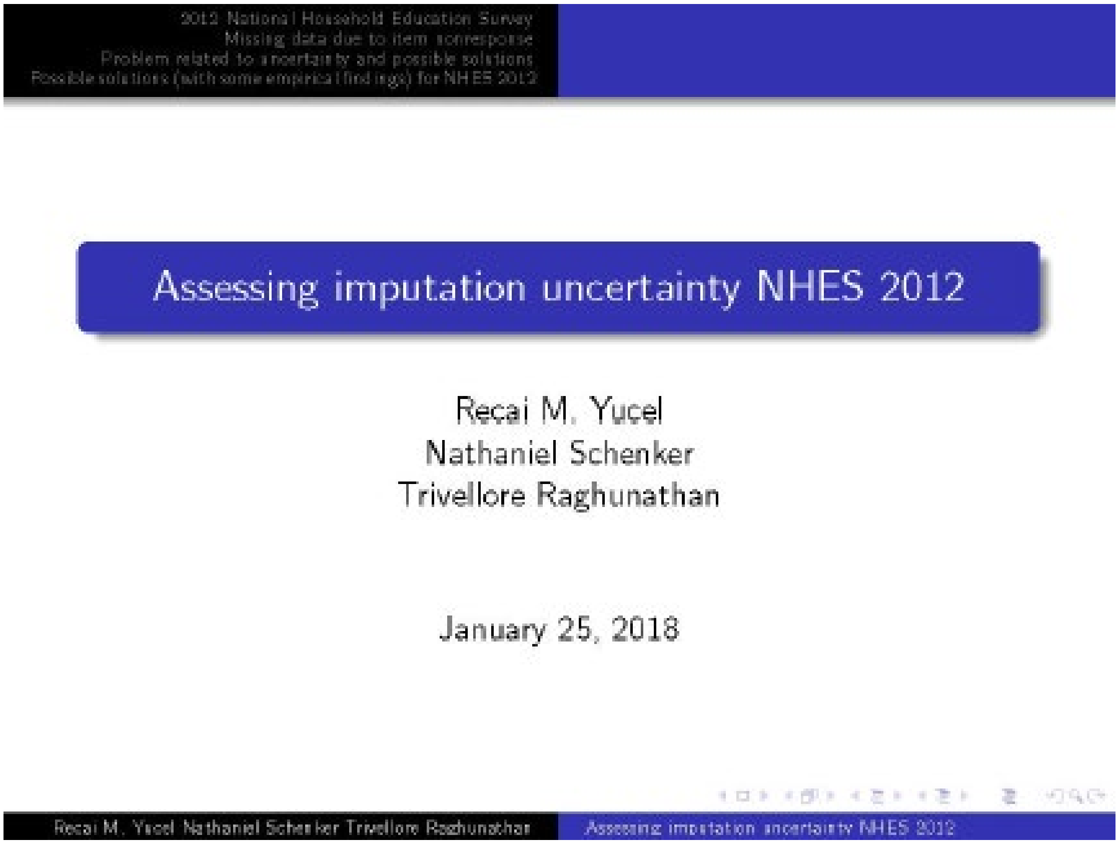
Table 5 shows that MI provides efficient point estimator than CC method but variance estimation is very conservative (more than 100% overestimation). Because of the serious positive bias of MI variance estimator, the statistical power of the test based on MI is actually lower than the CC method.



Conclusion

- We should understand the risks when MI is used in the production.
- MI has three main risks. Such risks should be clearly stated if we still want to use MI officially.
- Other options (such as fractional imputation) can also be considered.

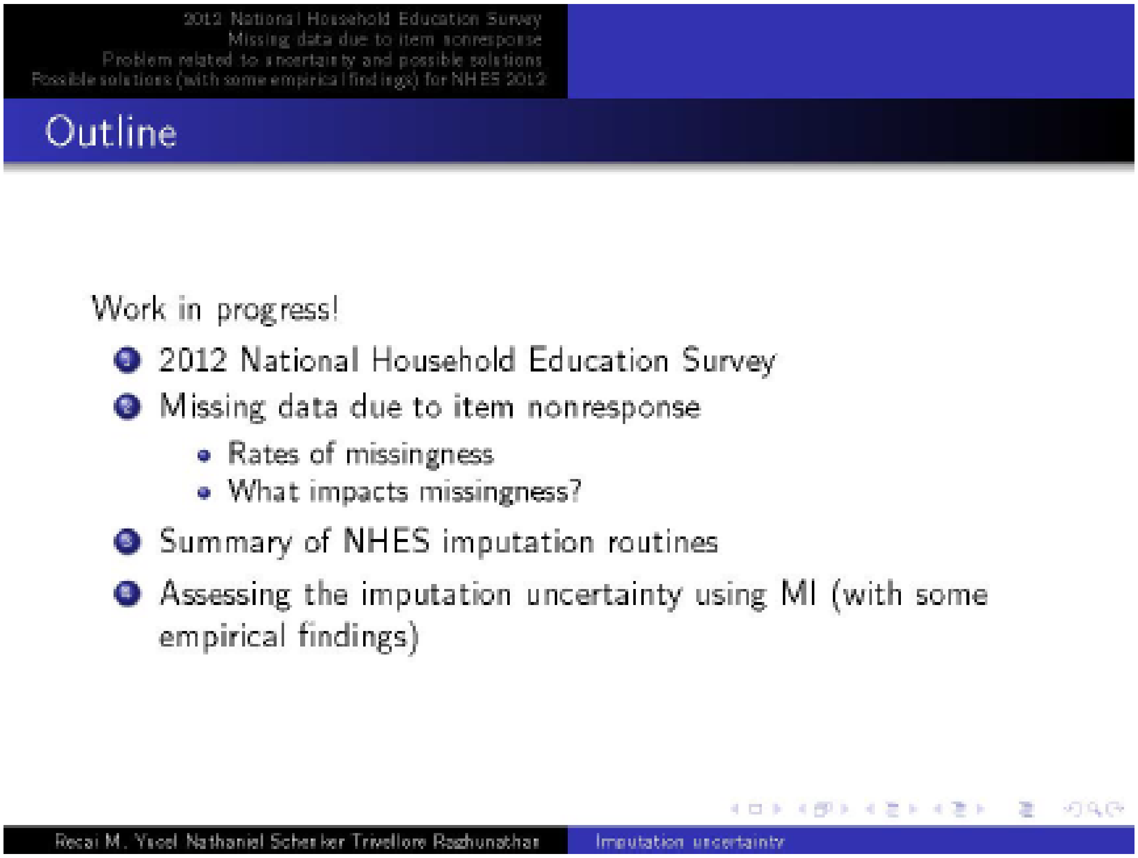
Assessing Imputation Uncertainty NHES 2012



Assessing imputation uncertainty NHES 2012

Recai M. Yucel
Nathaniel Schenker
Trivellore Raghunathan

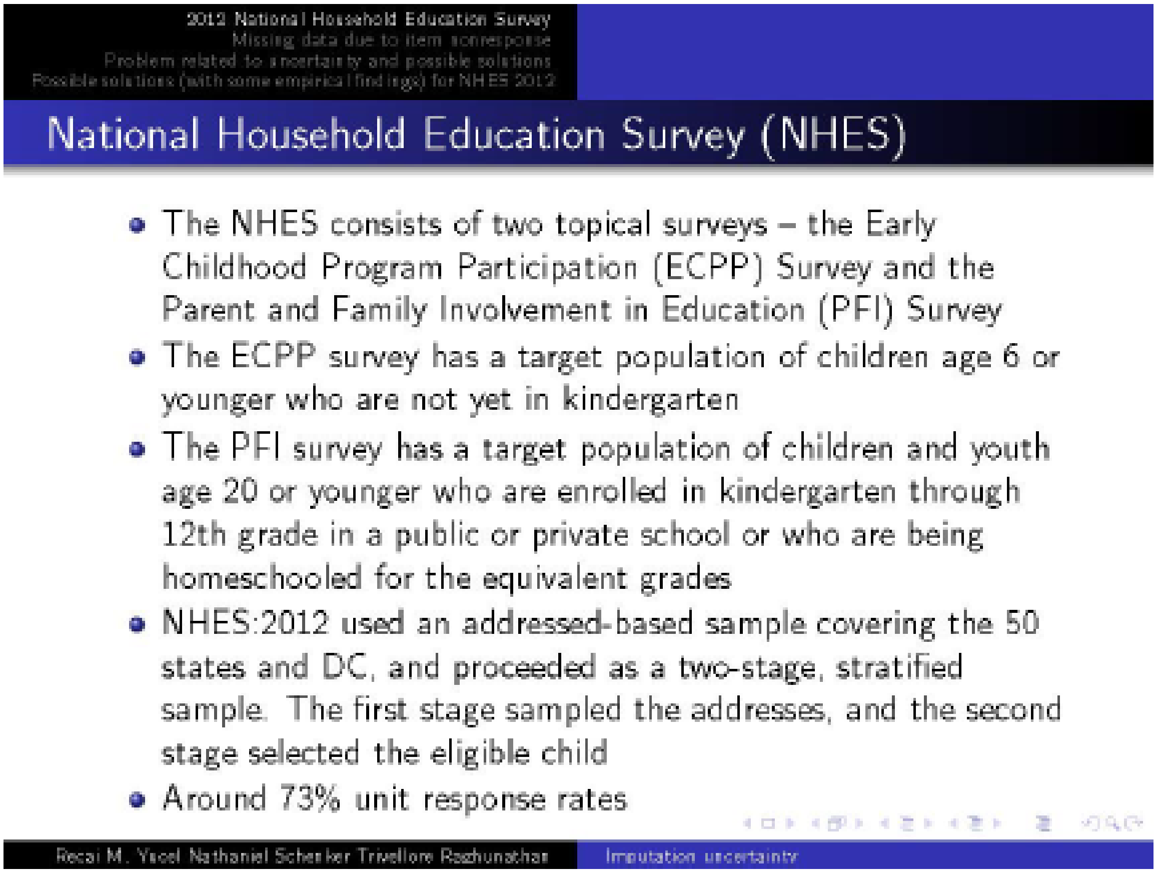
January 25, 2018



Outline

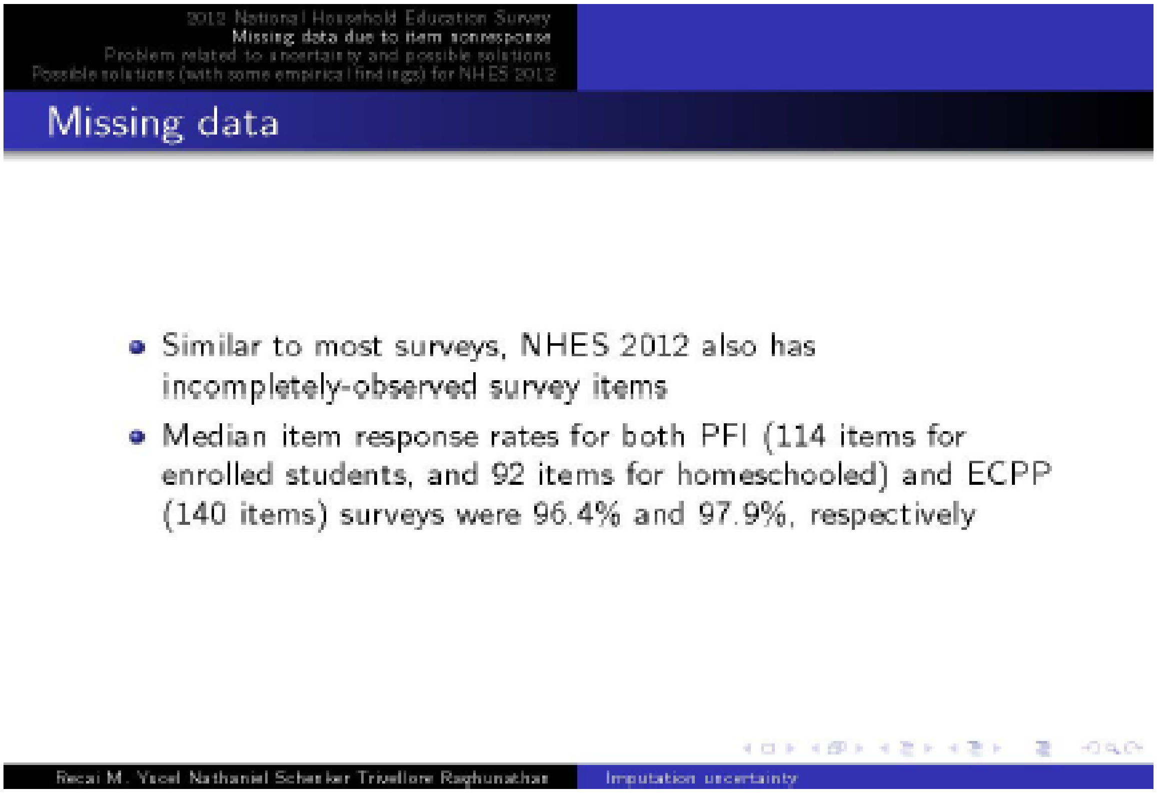
Work in progress!

1. 2012 National Household Education Survey
2. Missing data due to item nonresponse
   - Rates of missingness
   - What impacts missingness?
3. Summary of NHES imputation routines
4. Assessing the imputation uncertainty using MI (with some empirical findings)



National Household Education Survey (NHES)

- The NHES consists of two topical surveys – the Early Childhood Program Participation (ECPP) Survey and the Parent and Family Involvement in Education (PFI) Survey
- The ECPP survey has a target population of children age 6 or younger who are not yet in kindergarten
- The PFI survey has a target population of children and youth age 20 or younger who are enrolled in kindergarten through 12th grade in a public or private school or who are being homeschooled for the equivalent grades
- NHES:2012 used an addressed-based sample covering the 50 states and DC, and proceeded as a two-stage, stratified sample. The first stage sampled the addresses, and the second stage selected the eligible child
- Around 73% unit response rates



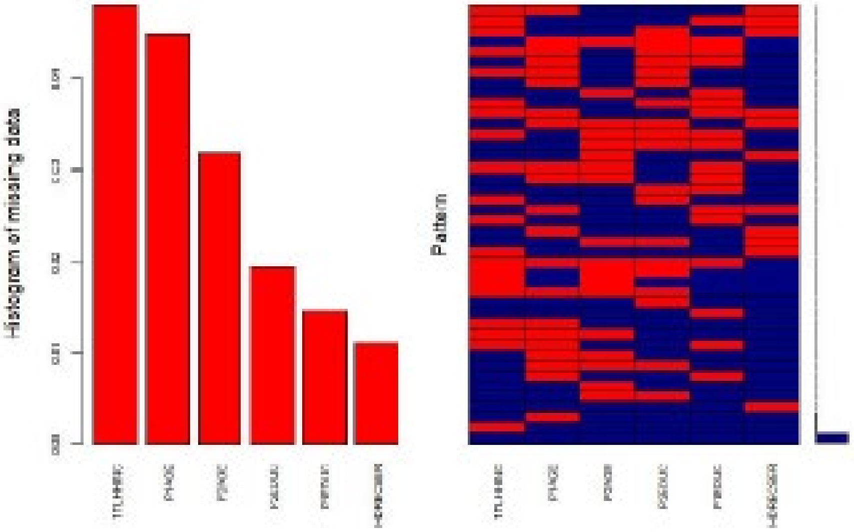
Missing data

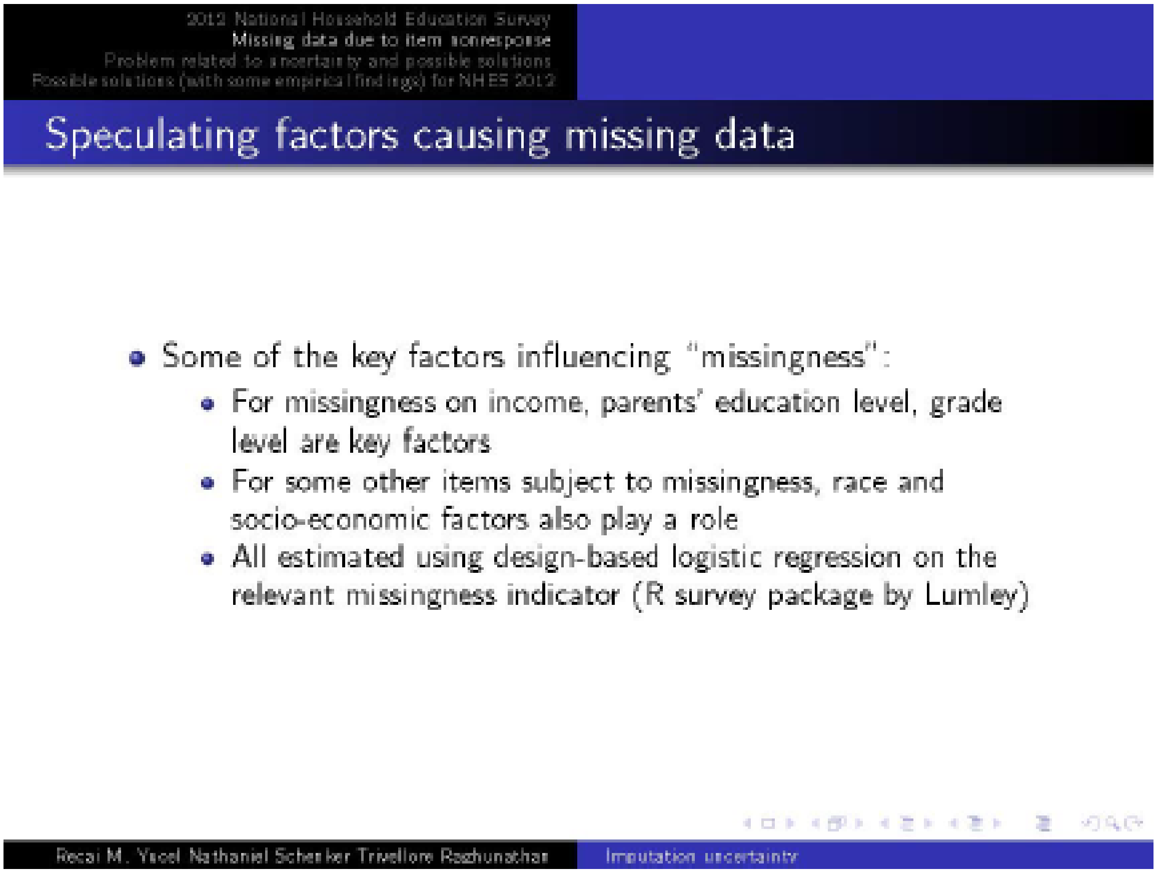
- Similar to most surveys, NHES 2012 also has incompletely-observed survey items
- Median item response rates for both PFI (114 items for enrolled students, and 92 items for homeschooled) and ECPP (140 items) surveys were 96.4% and 97.9%, respectively

Missing data: example

- For this presentation, consider an analysis involving six variables: Age (783 out of 17563 respondents in PFI), Education (256 parent 1, 344 parent 2), Total Household Income (846 missing out of 17563) and indicator for receiving special health services
- 15663 cases from PFI module have complete data (1900 cases have at least one item missing) in this subset of variables

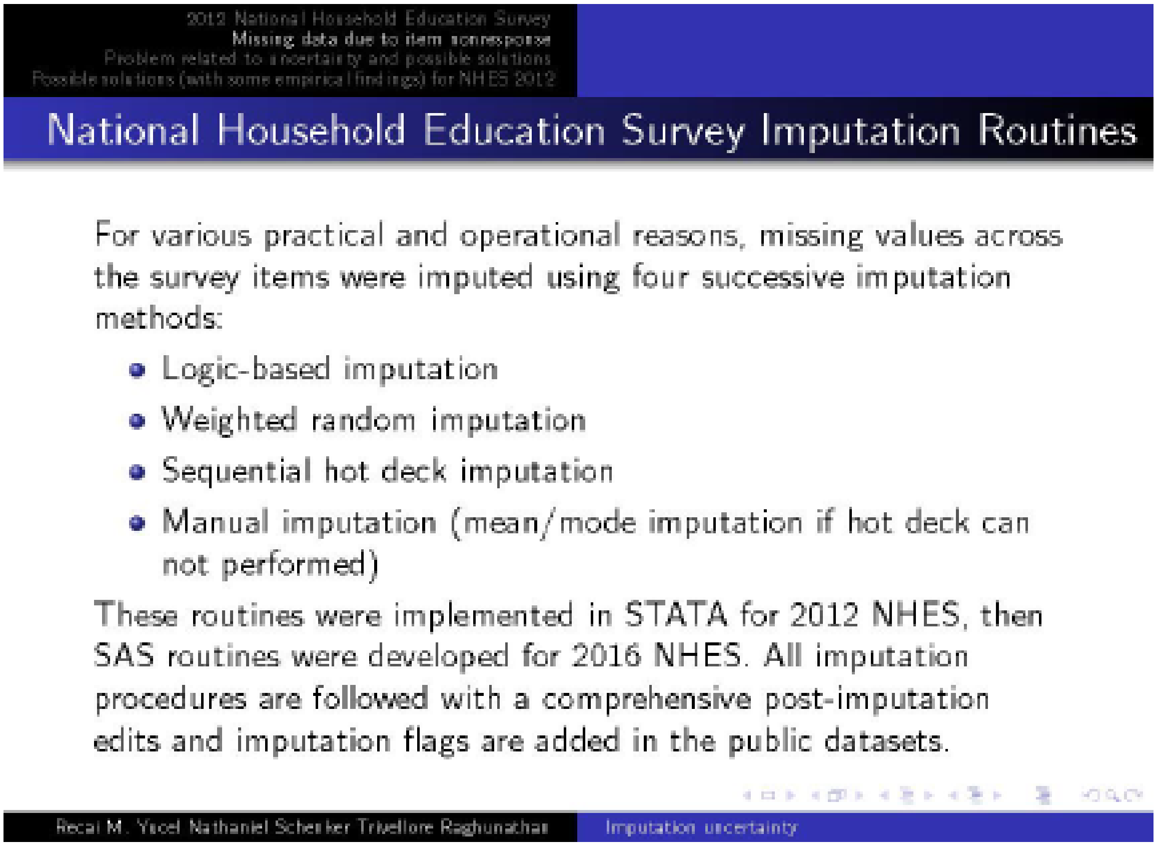
Example missing data pattern





Speculating factors causing missing data

- Some of the key factors influencing "missingness":
  - For missingness on income, parents' education level, grade level are key factors
  - For some other items subject to missingness, race and socio-economic factors also play a role
  - All estimated using design-based logistic regression on the relevant missingness indicator (R survey package by Lumley)

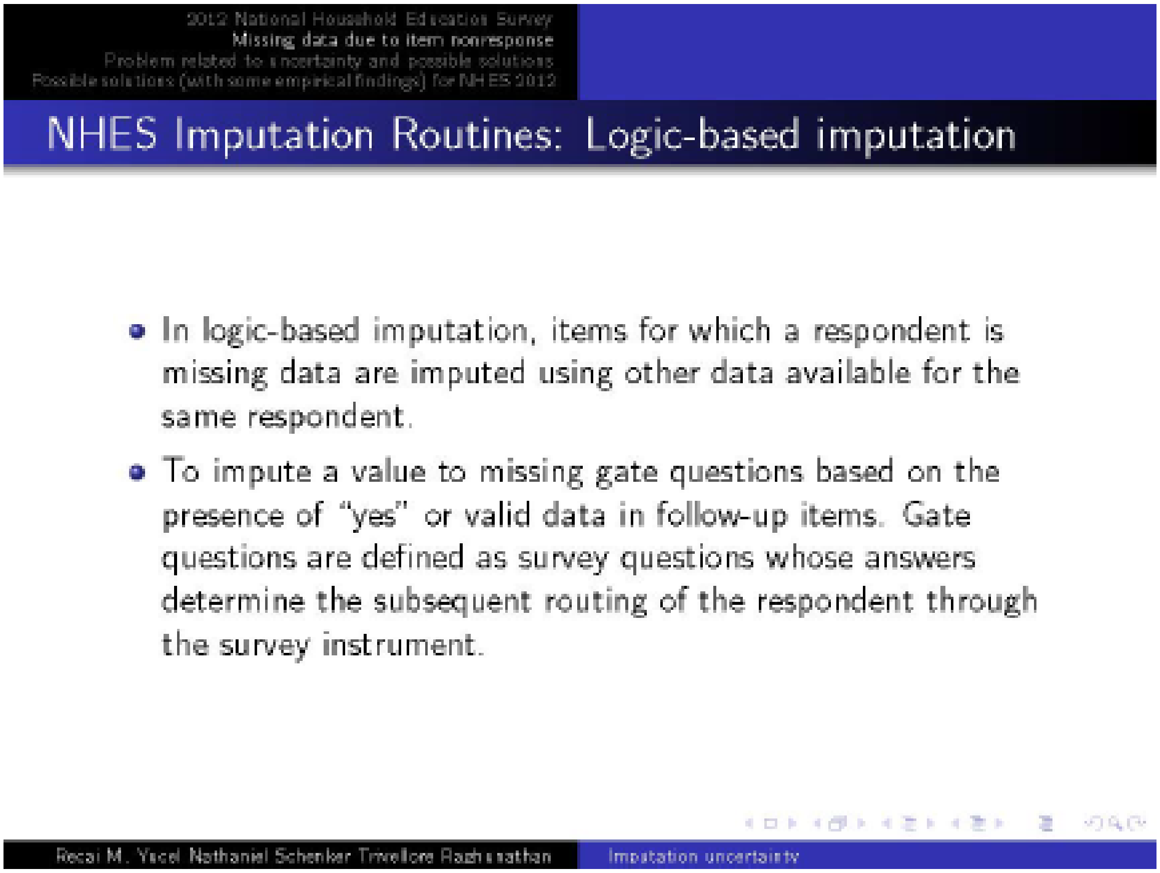


National Household Education Survey Imputation Routines

For various practical and operational reasons, missing values across the survey items were imputed using four successive imputation methods:

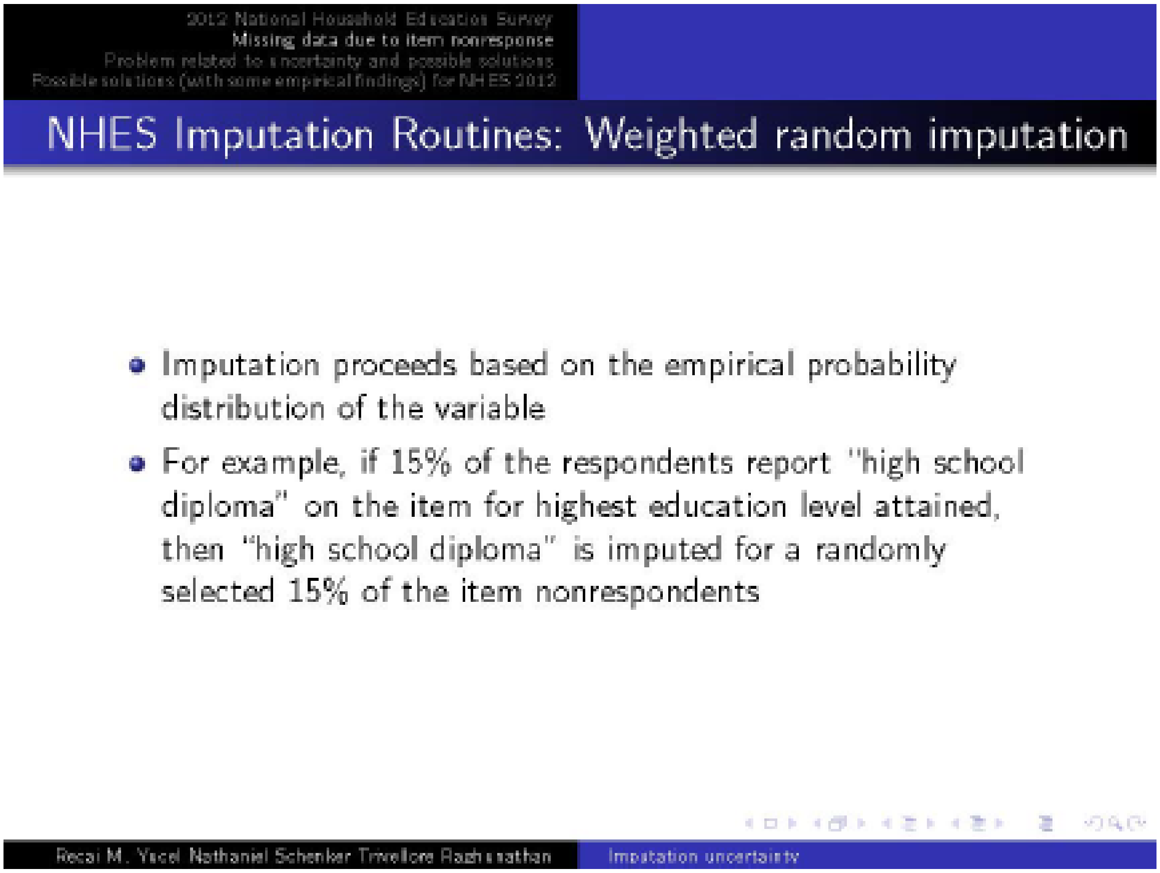
- Logic-based imputation
- Weighted random imputation
- Sequential hot deck imputation
- Manual imputation (mean/mode imputation if hot deck can not performed)

These routines were implemented in STATA for 2012 NHES, then SAS routines were developed for 2016 NHES. All imputation procedures are followed with a comprehensive post-imputation edits and imputation flags are added in the public datasets.



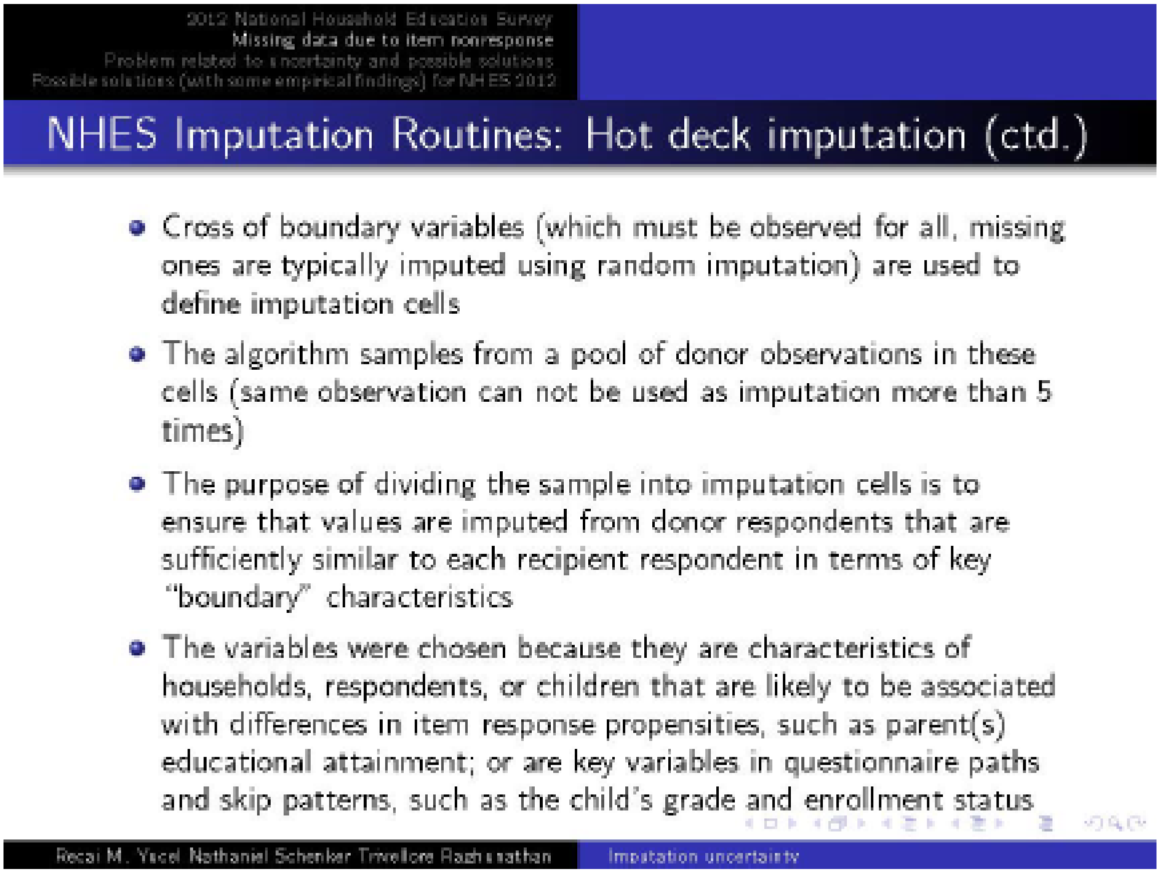
NHES Imputation Routines: Logic-based imputation

- In logic-based imputation, items for which a respondent is missing data are imputed using other data available for the same respondent.
- To impute a value to missing gate questions based on the presence of "yes" or valid data in follow-up items. Gate questions are defined as survey questions whose answers determine the subsequent routing of the respondent through the survey instrument.



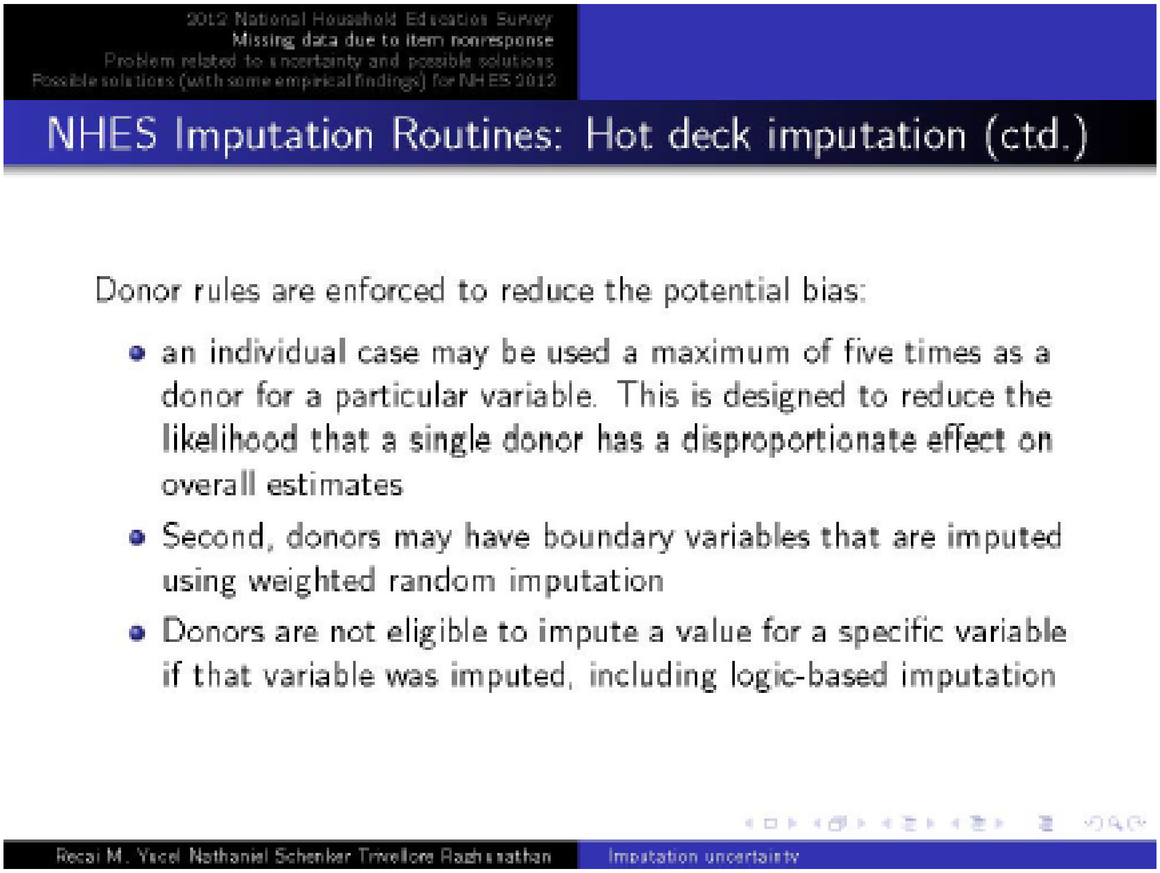
NHES Imputation Routines: Weighted random imputation

- Imputation proceeds based on the empirical probability distribution of the variable
- For example, if 15% of the respondents report "high school diploma" on the item for highest education level attained, then "high school diploma" is imputed for a randomly selected 15% of the item nonrespondents



NHES Imputation Routines: Hot deck imputation (ctd.)

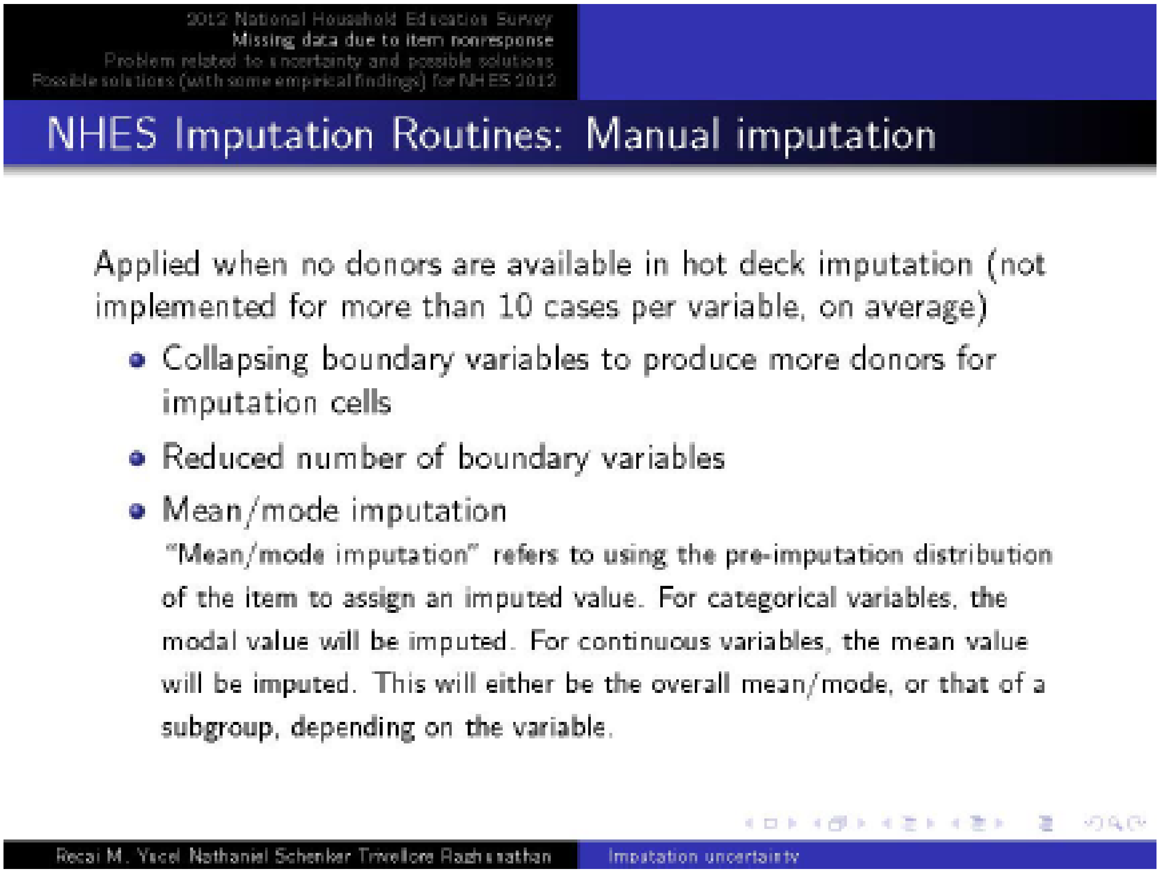
- Cross of boundary variables (which must be observed for all, missing ones are typically imputed using random imputation) are used to define imputation cells
- The algorithm samples from a pool of donor observations in these cells (same observation can not be used as imputation more than 5 times)
- The purpose of dividing the sample into imputation cells is to ensure that values are imputed from donor respondents that are sufficiently similar to each recipient respondent in terms of key "boundary" characteristics
- The variables were chosen because they are characteristics of households, respondents, or children that are likely to be associated with differences in item response propensities, such as parent(s) educational attainment; or are key variables in questionnaire paths and skip patterns, such as the child's grade and enrollment status



NHES Imputation Routines: Hot deck imputation (ctd.)

Donor rules are enforced to reduce the potential bias:

- an individual case may be used a maximum of five times as a donor for a particular variable. This is designed to reduce the likelihood that a single donor has a disproportionate effect on overall estimates
- Second, donors may have boundary variables that are imputed using weighted random imputation
- Donors are not eligible to impute a value for a specific variable if that variable was imputed, including logic-based imputation

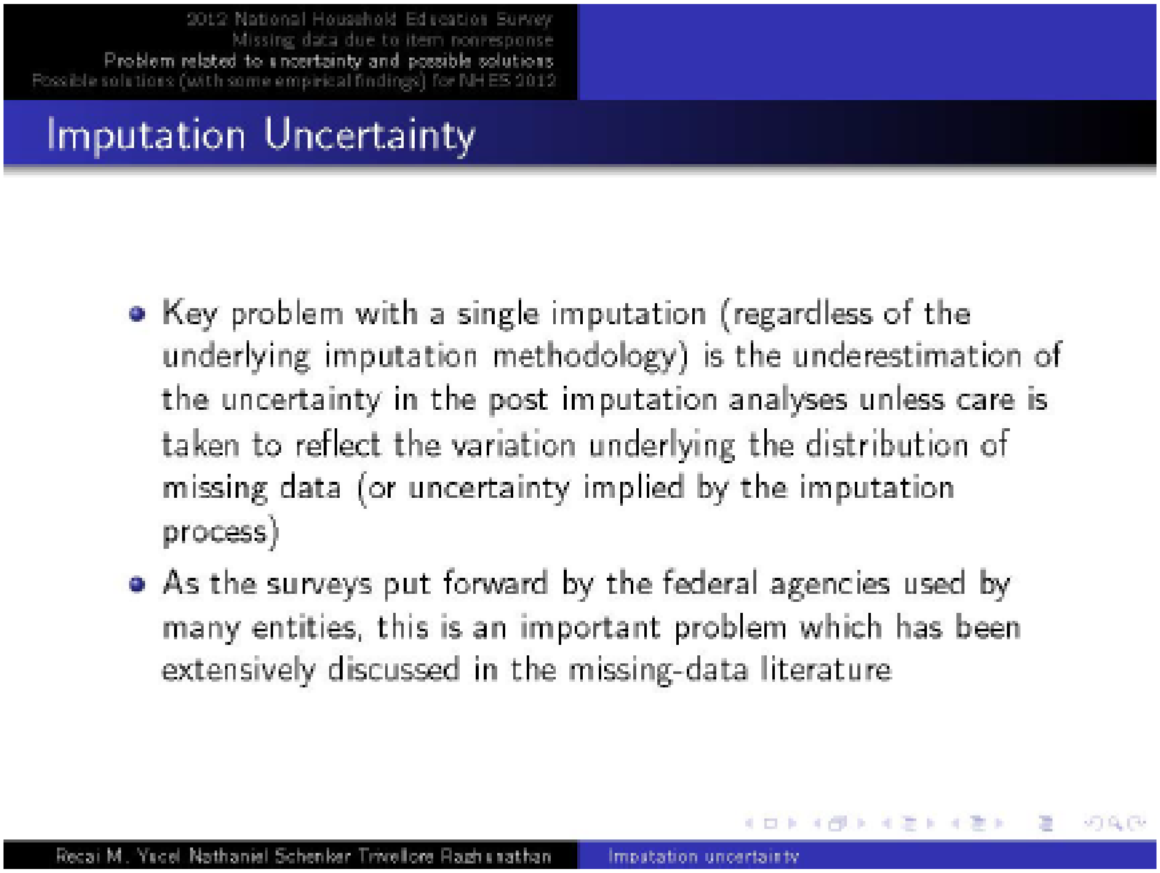


NHES Imputation Routines: Manual imputation

Applied when no donors are available in hot deck imputation (not implemented for more than 10 cases per variable, on average)

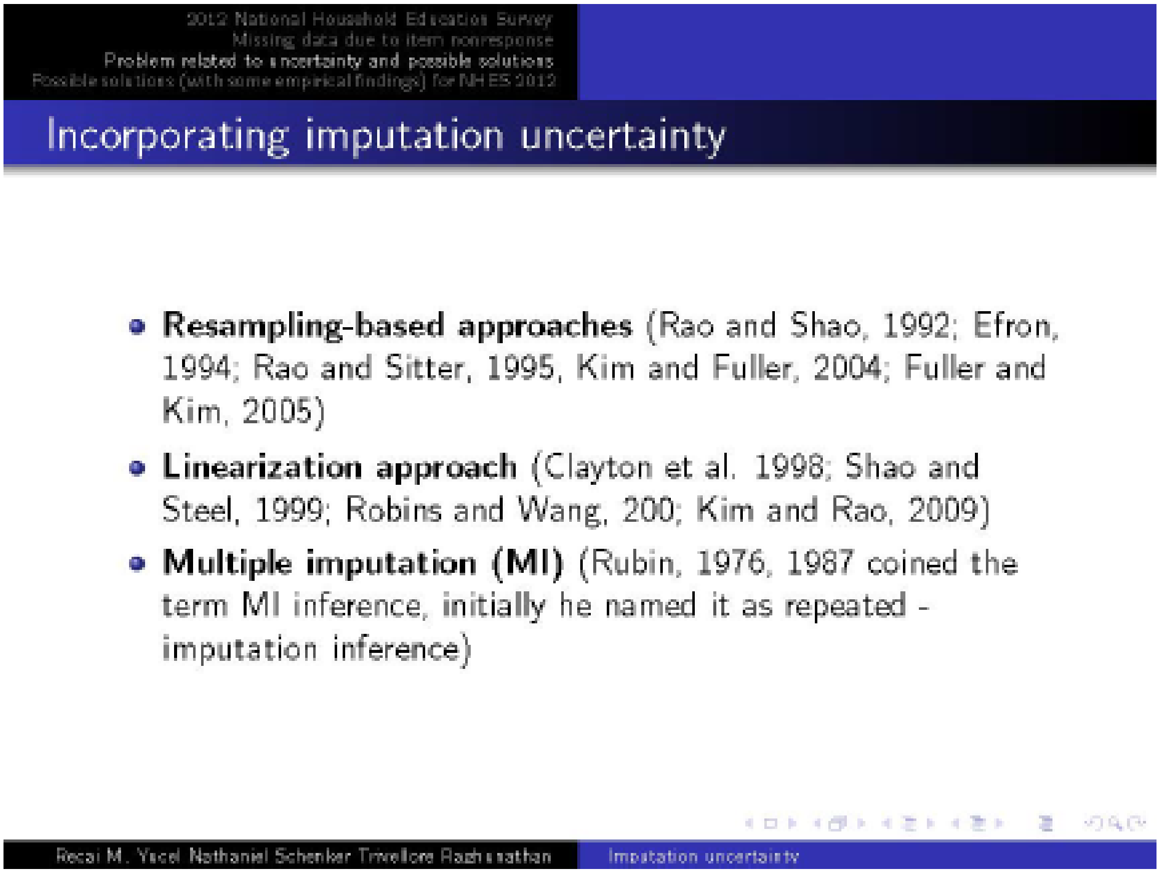
- Collapsing boundary variables to produce more donors for imputation cells
- Reduced number of boundary variables
- Mean/mode imputation

  "Mean/mode imputation" refers to using the pre-imputation distribution of the item to assign an imputed value. For categorical variables, the modal value will be imputed. For continuous variables, the mean value will be imputed. This will either be the overall mean/mode, or that of a subgroup, depending on the variable.



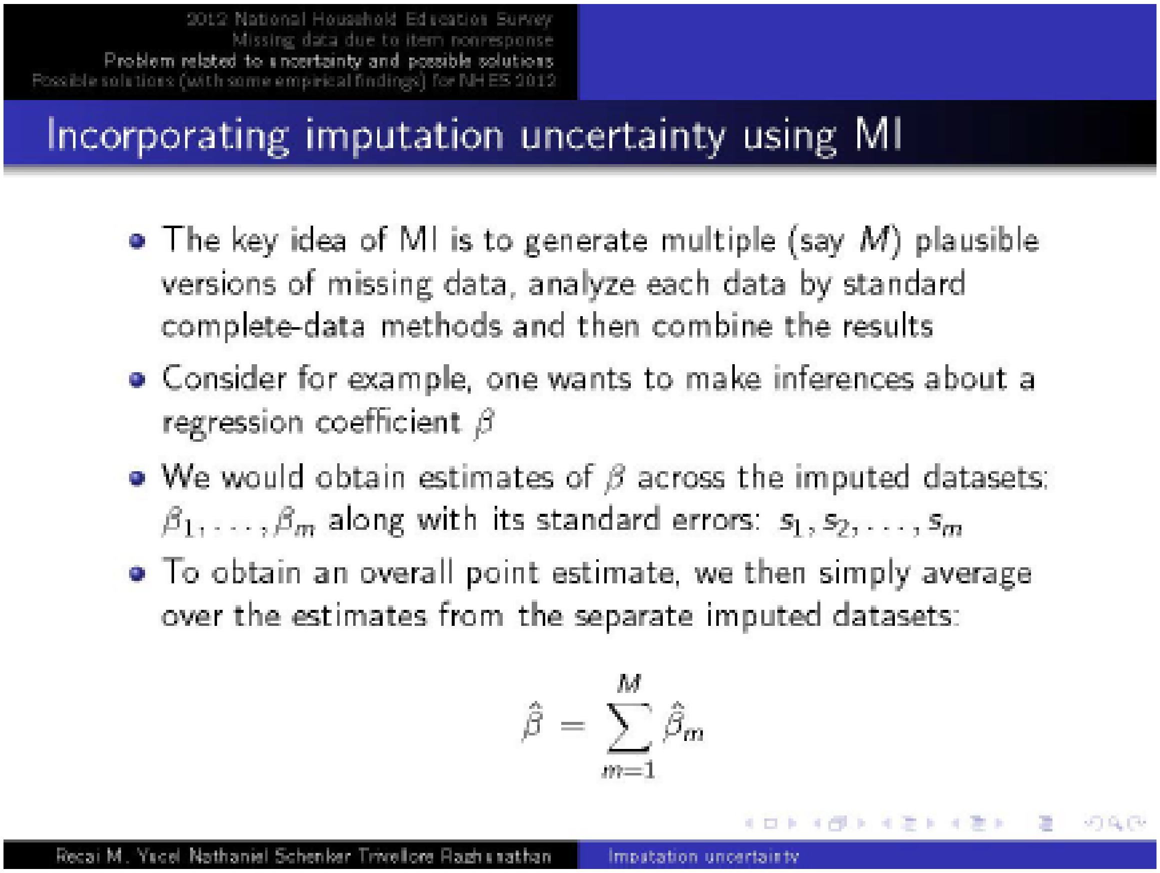
Imputation Uncertainty

- Key problem with a single imputation (regardless of the underlying imputation methodology) is the underestimation of the uncertainty in the post imputation analyses unless care is taken to reflect the variation underlying the distribution of missing data (or uncertainty implied by the imputation process)
- As the surveys put forward by the federal agencies used by many entities, this is an important problem which has been extensively discussed in the missing-data literature



Incorporating imputation uncertainty

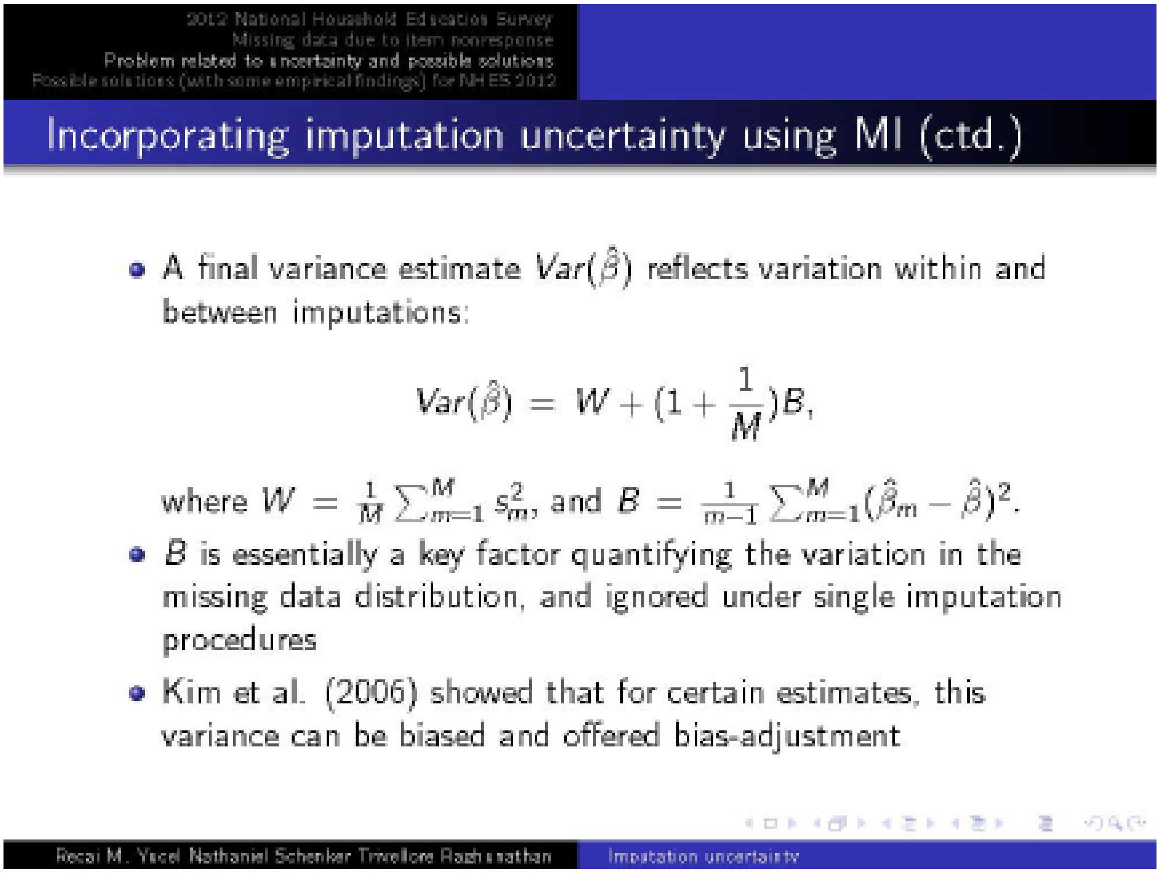
- **Resampling-based approaches** (Rao and Shao, 1992; Efron, 1994; Rao and Sitter, 1995, Kim and Fuller, 2004; Fuller and Kim, 2005)
- **Linearization approach** (Clayton et al. 1998; Shao and Steel, 1999; Robins and Wang, 200; Kim and Rao, 2009)
- **Multiple imputation (MI)** (Rubin, 1976, 1987 coined the term MI inference, initially he named it as repeated - imputation inference)



Incorporating imputation uncertainty using MI

- The key idea of MI is to generate multiple (say M) plausible versions of missing data, analyze each data by standard complete-data methods and then combine the results
- Consider for example, one wants to make inferences about a regression coefficient β
- We would obtain estimates of β across the imputed datasets: $\beta_1, \ldots, \beta_m$ along with its standard errors: $s_1, s_2, \ldots, s_m$
- To obtain an overall point estimate, we then simply average over the estimates from the separate imputed datasets:

$$\hat{\beta} = \sum_{m=1}^{M} \hat{\beta}_m$$



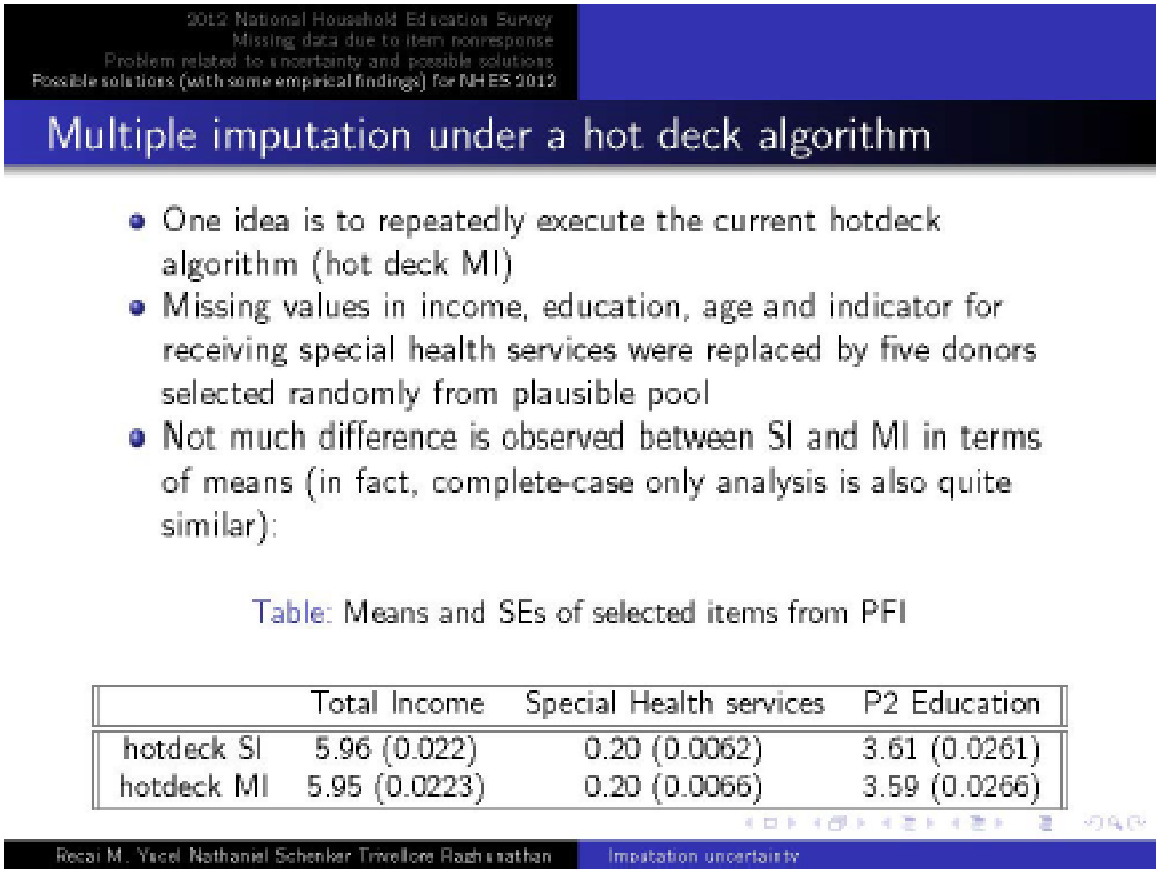
Incorporating imputation uncertainty using MI (ctd.)

- A final variance estimate $Var(\hat{\beta})$ reflects variation within and between imputations:

$$Var(\hat{\beta}) = W + (1 + \frac{1}{M})B,$$

where $W = \frac{1}{M}\sum_{m=1}^{M} s_m^2$, and $B = \frac{1}{m-1}\sum_{m=1}^{M}(\hat{\beta}_m - \hat{\beta})^2$.

- B is essentially a key factor quantifying the variation in the missing data distribution, and ignored under single imputation procedures
- Kim et al. (2006) showed that for certain estimates, this variance can be biased and offered bias-adjustment

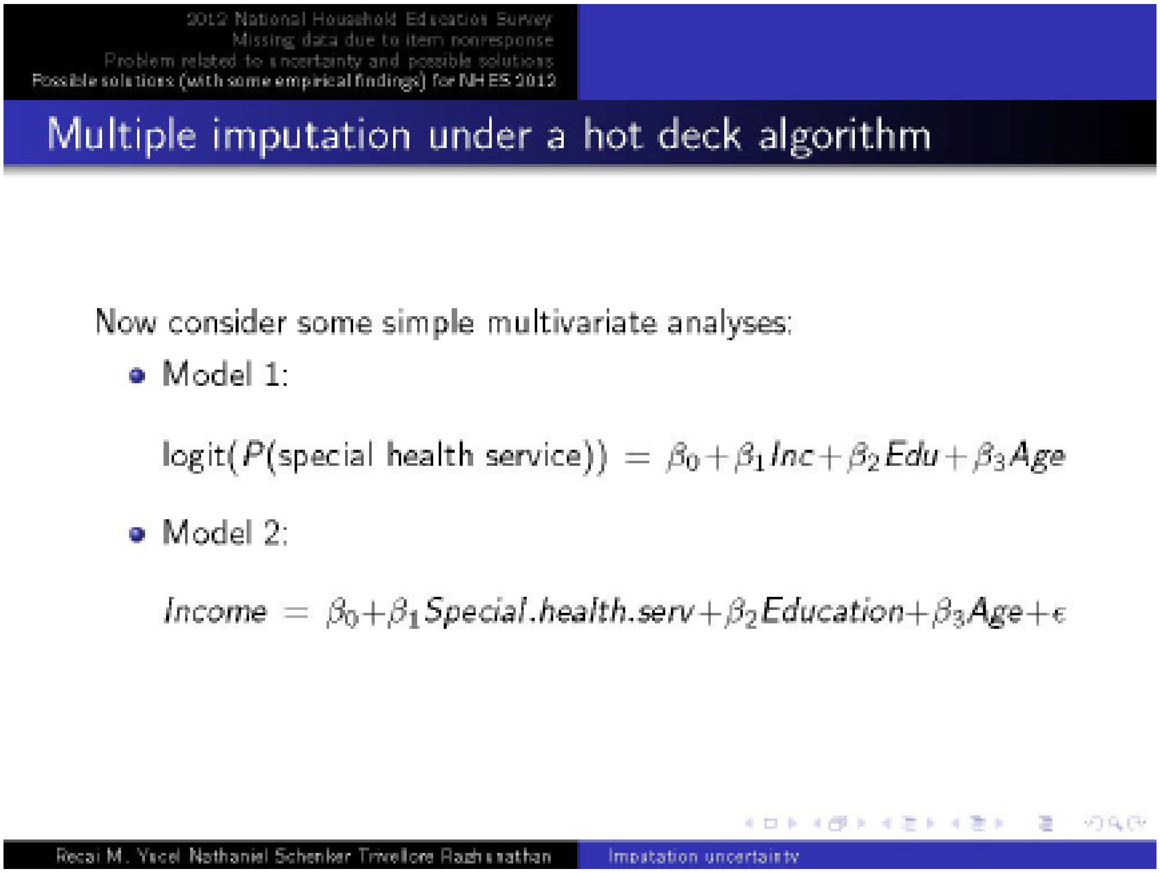


Multiple imputation under a hot deck algorithm

- One idea is to repeatedly execute the current hotdeck algorithm (hot deck MI)
- Missing values in income, education, age and indicator for receiving special health services were replaced by five donors selected randomly from plausible pool
- Not much difference is observed between SI and MI in terms of means (in fact, complete-case only analysis is also quite similar):

Table: Means and SEs of selected items from PFI

	Total Income	Special Health services	P2 Education
hotdeck SI	5.96 (0.022)	0.20 (0.0062)	3.61 (0.0261)
hotdeck MI	5.95 (0.0223)	0.20 (0.0066)	3.59 (0.0266)



Multiple imputation under a hot deck algorithm

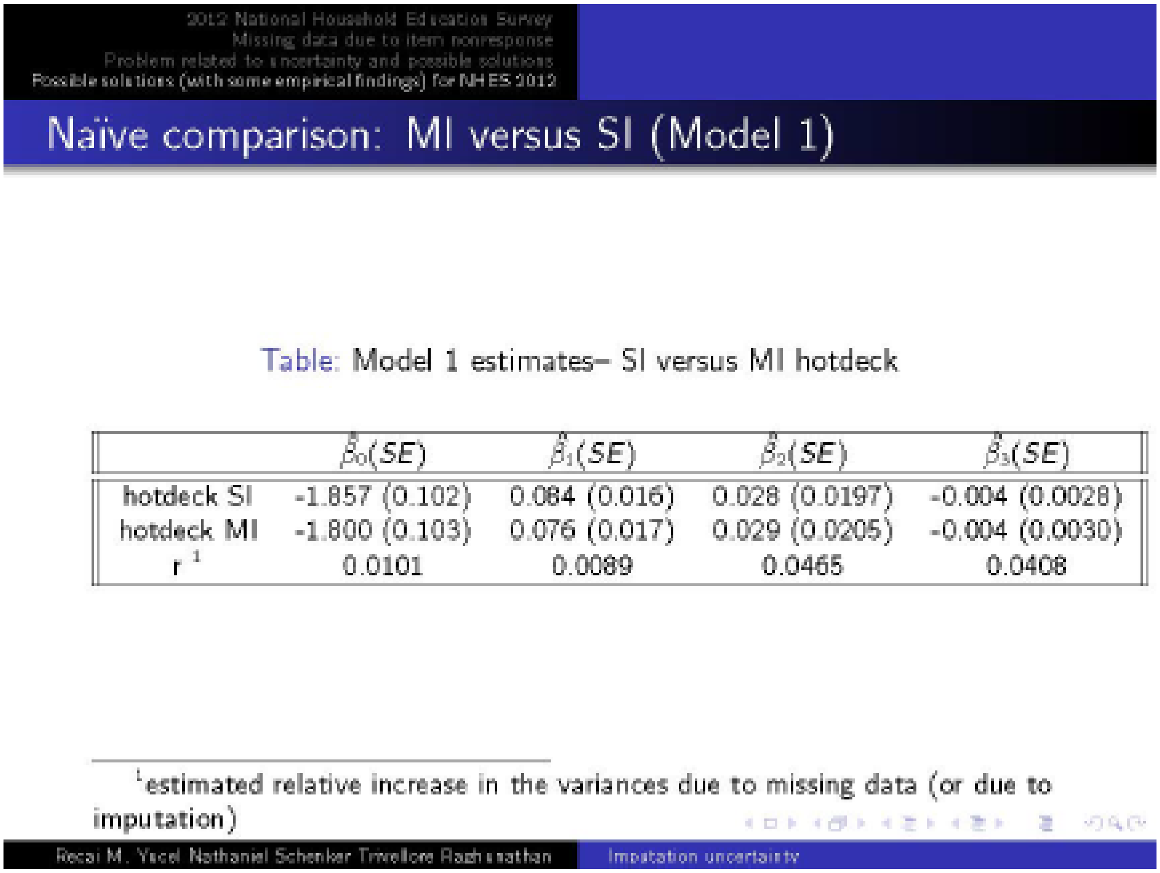
Now consider some simple multivariate analyses:

- Model 1:

$$\text{logit}(P(\text{special health service})) = \beta_0 + \beta_1 Inc + \beta_2 Edu + \beta_3 Age$$

- Model 2:

$$Income = \beta_0 + \beta_1 Special.health.serv + \beta_2 Education + \beta_3 Age + \epsilon$$

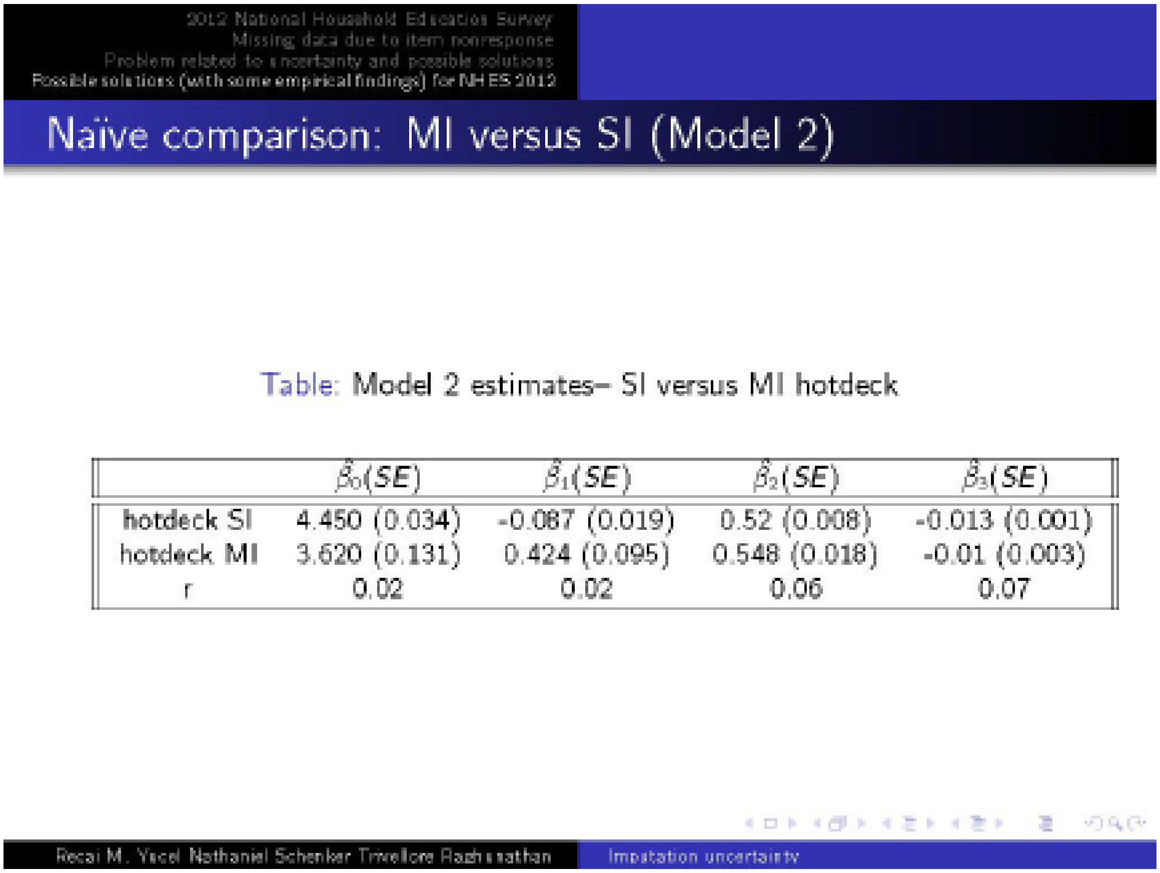


Naïve comparison: MI versus SI (Model 1)

Table: Model 1 estimates– SI versus MI hotdeck

	$\hat{\beta}_0(SE)$	$\hat{\beta}_1(SE)$	$\hat{\beta}_2(SE)$	$\hat{\beta}_3(SE)$
hotdeck SI	-1.857 (0.102)	0.084 (0.016)	0.028 (0.0197)	-0.004 (0.0028)
hotdeck MI	-1.800 (0.103)	0.076 (0.017)	0.029 (0.0205)	-0.004 (0.0030)
r [1]	0.0101	0.0089	0.0465	0.0408

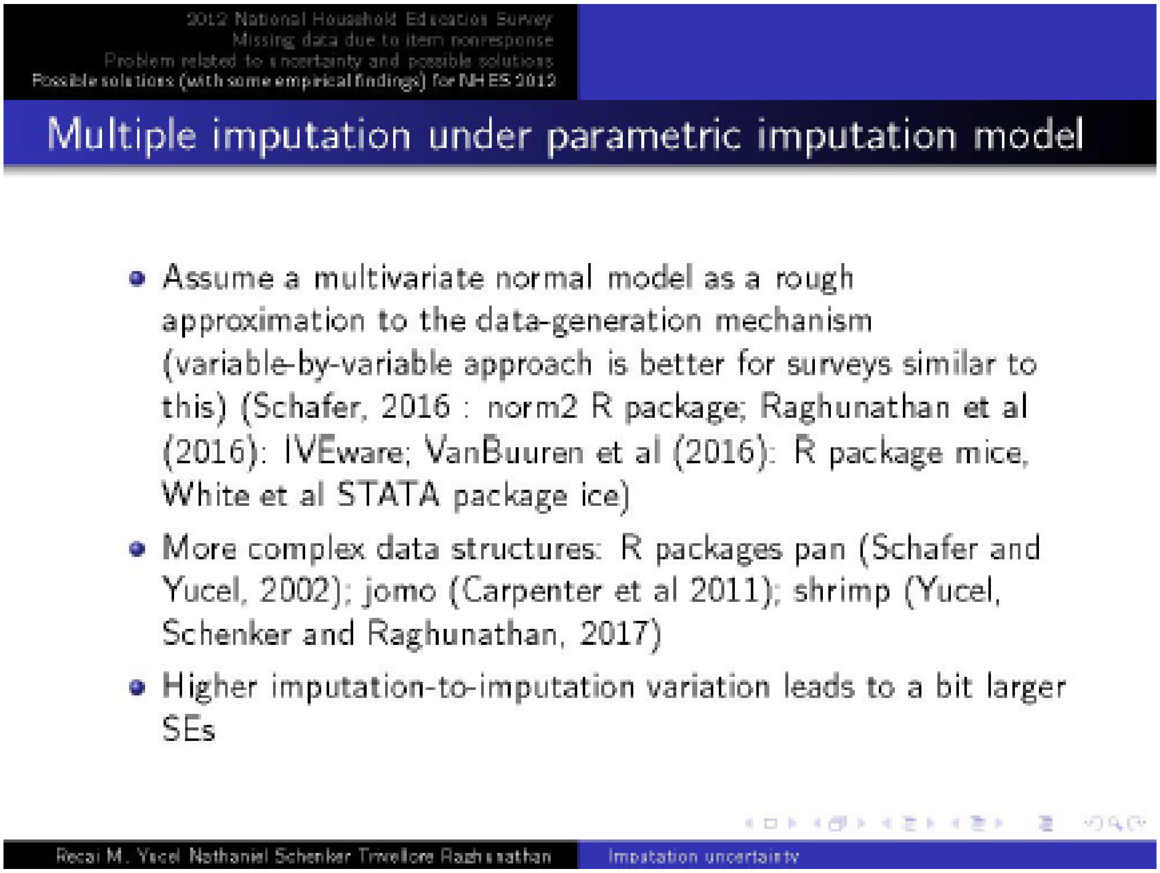
[1] estimated relative increase in the variances due to missing data (or due to imputation)



Naïve comparison: MI versus SI (Model 2)

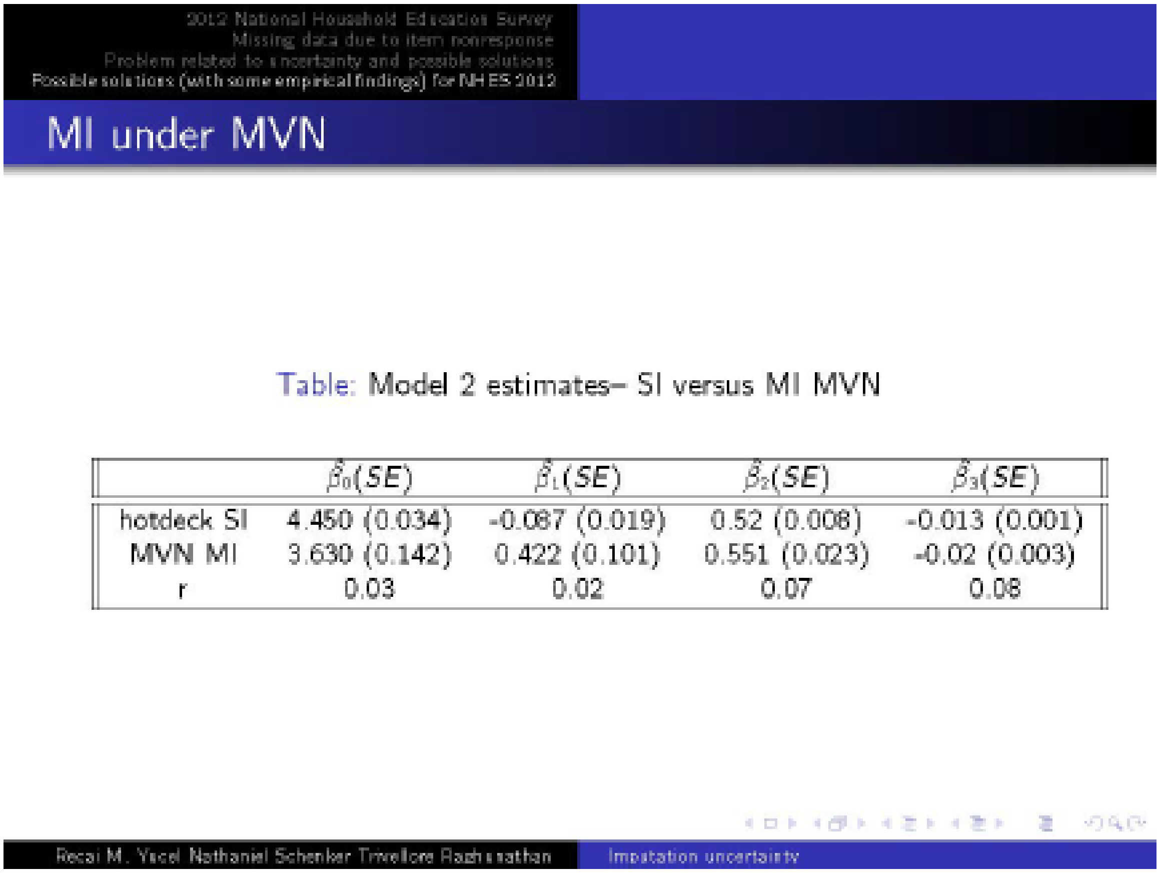
Table: Model 2 estimates– SI versus MI hotdeck

	$\hat{\beta}_0(SE)$	$\hat{\beta}_1(SE)$	$\hat{\beta}_2(SE)$	$\hat{\beta}_3(SE)$
hotdeck SI	4.450 (0.034)	-0.087 (0.019)	0.52 (0.008)	-0.013 (0.001)
hotdeck MI	3.620 (0.131)	0.424 (0.095)	0.548 (0.018)	-0.01 (0.003)
r	0.02	0.02	0.06	0.07



Multiple imputation under parametric imputation model

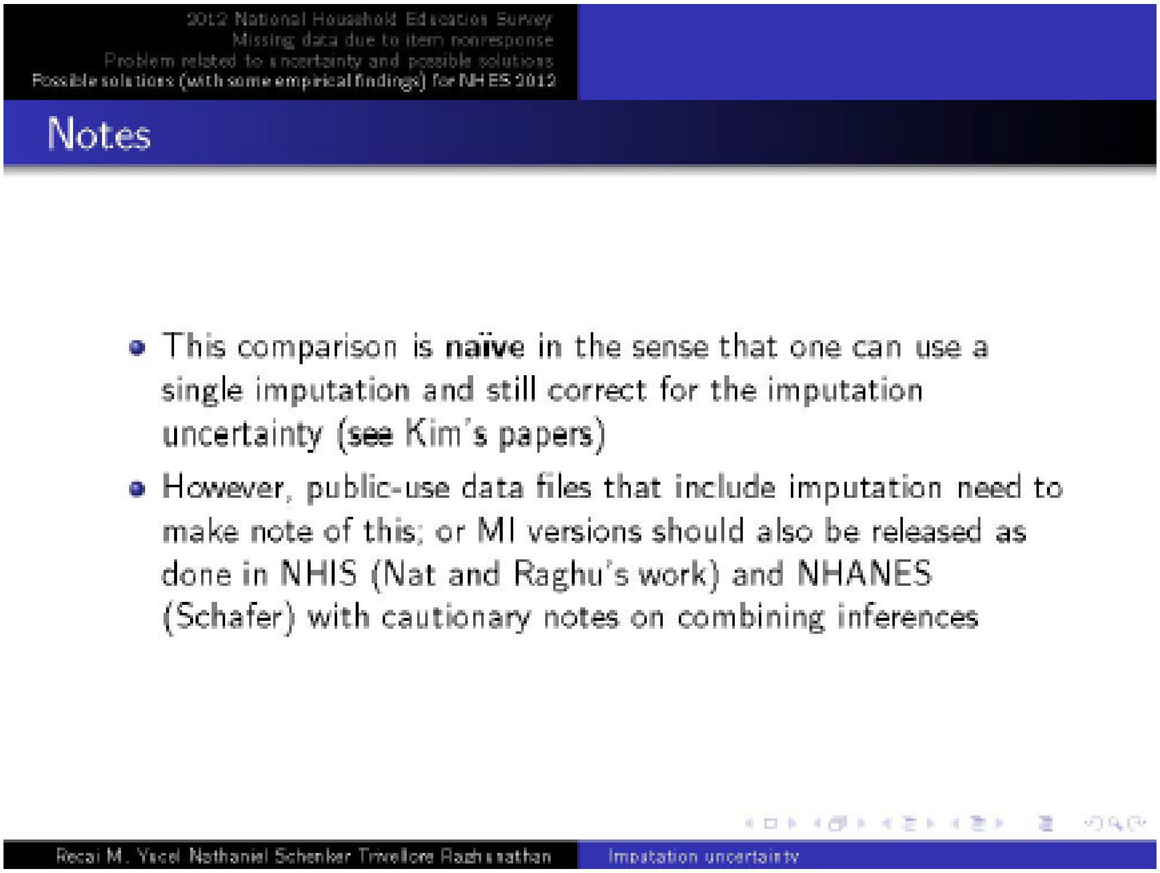
- Assume a multivariate normal model as a rough approximation to the data-generation mechanism (variable-by-variable approach is better for surveys similar to this) (Schafer, 2016 : norm2 R package; Raghunathan et al (2016): IVEware; VanBuuren et al (2016): R package mice, White et al STATA package ice)
- More complex data structures: R packages pan (Schafer and Yucel, 2002); jomo (Carpenter et al 2011); shrimp (Yucel, Schenker and Raghunathan, 2017)
- Higher imputation-to-imputation variation leads to a bit larger SEs



MI under MVN

Table: Model 2 estimates– SI versus MI MVN

	$\hat{\beta}_0(SE)$	$\hat{\beta}_1(SE)$	$\hat{\beta}_2(SE)$	$\hat{\beta}_3(SE)$
hotdeck SI	4.450 (0.034)	-0.087 (0.019)	0.52 (0.008)	-0.013 (0.001)
MVN MI	3.630 (0.142)	0.422 (0.101)	0.551 (0.023)	-0.02 (0.003)
r	0.03	0.02	0.07	0.08



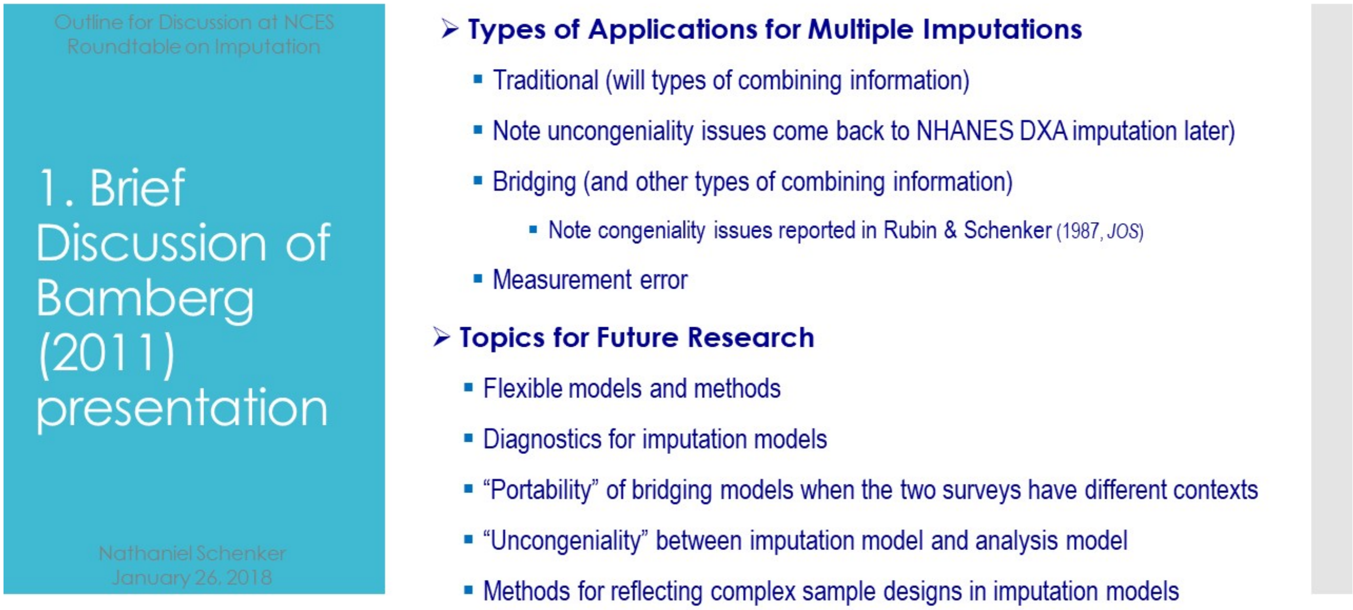
Notes

- This comparison is **naïve** in the sense that one can use a single imputation and still correct for the imputation uncertainty (see Kim's papers)
- However, public-use data files that include imputation need to make note of this; or MI versions should also be released as done in NHIS (Nat and Raghu's work) and NHANES (Schafer) with cautionary notes on combining inferences

Outline for Discussion at NCES Roundtable on Imputation

Outline for Discussion at NCES Roundtable on Imputation

Nathaniel Schenker
January 26, 2018



Outline for Discussion at NCES Roundtable on Imputation

1. Brief Discussion of Bamberg (2011) presentation

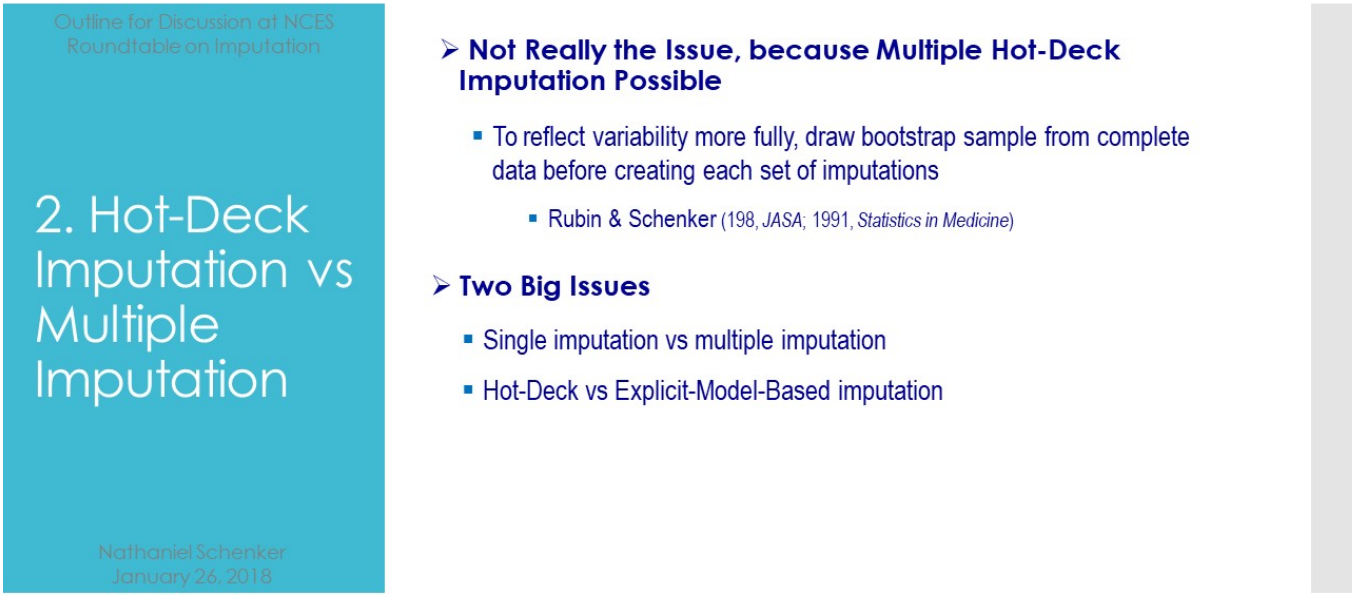
Nathaniel Schenker
January 26, 2018

➢ **Types of Applications for Multiple Imputations**

- Traditional (will types of combining information)
- Note uncongeniality issues come back to NHANES DXA imputation later)
- Bridging (and other types of combining information)
  - Note congeniality issues reported in Rubin & Schenker (1987, *JOS*)
- Measurement error

➢ **Topics for Future Research**

- Flexible models and methods
- Diagnostics for imputation models
- "Portability" of bridging models when the two surveys have different contexts
- "Uncongeniality" between imputation model and analysis model
- Methods for reflecting complex sample designs in imputation models

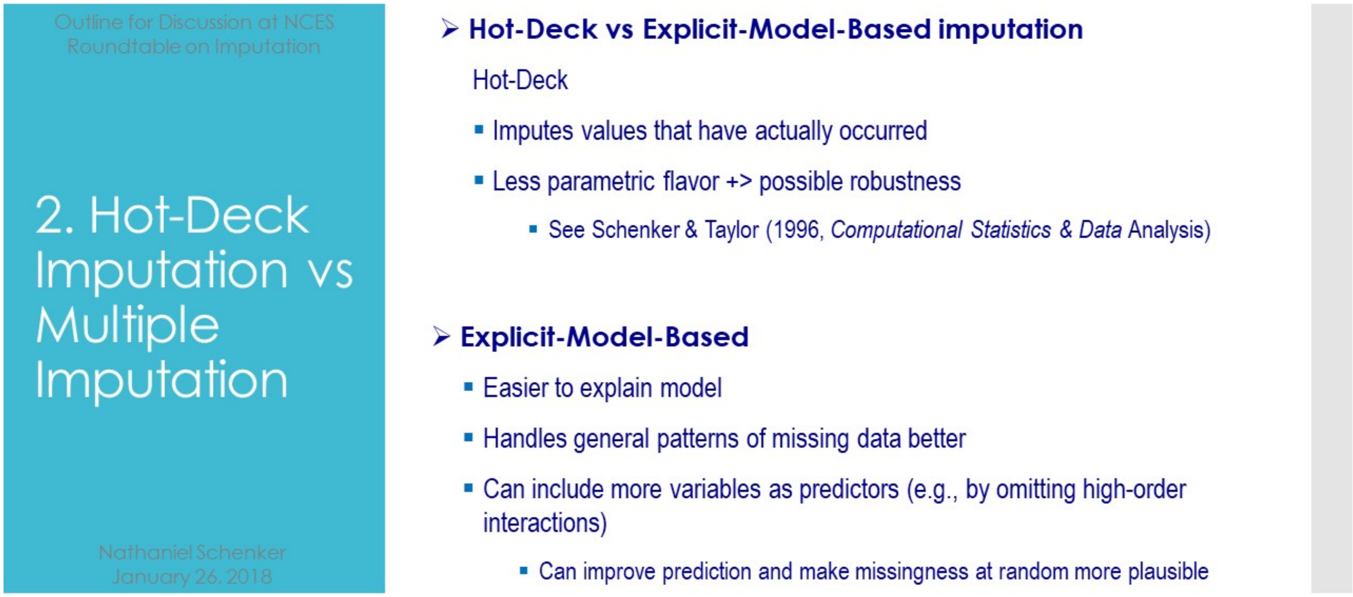


Outline for Discussion at NCES Roundtable on Imputation

2. Hot-Deck Imputation vs Multiple Imputation

- **Not Really the Issue, because Multiple Hot-Deck Imputation Possible**
  - To reflect variability more fully, draw bootstrap sample from complete data before creating each set of imputations
    - Rubin & Schenker (198, *JASA*; 1991, *Statistics in Medicine*)
- **Two Big Issues**
  - Single imputation vs multiple imputation
  - Hot-Deck vs Explicit-Model-Based imputation

Nathaniel Schenker
January 26, 2018



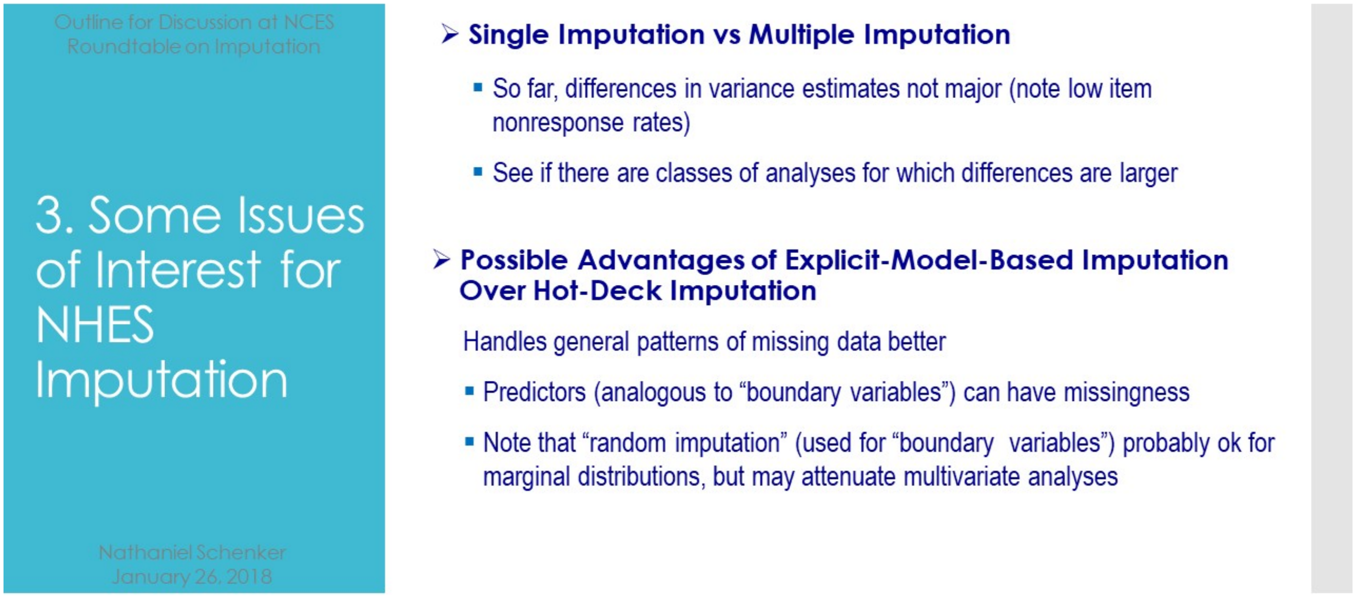
Outline for Discussion at NCES Roundtable on Imputation

2. Hot-Deck Imputation vs Multiple Imputation

- **Hot-Deck vs Explicit-Model-Based imputation**

  Hot-Deck
  - Imputes values that have actually occurred
  - Less parametric flavor +> possible robustness
    - See Schenker & Taylor (1996, *Computational Statistics & Data* Analysis)
- **Explicit-Model-Based**
  - Easier to explain model
  - Handles general patterns of missing data better
  - Can include more variables as predictors (e.g., by omitting high-order interactions)
    - Can improve prediction and make missingness at random more plausible

Nathaniel Schenker
January 26, 2018



Outline for Discussion at NCES Roundtable on Imputation

3. Some Issues of Interest for NHES Imputation

Nathaniel Schenker
January 26, 2018

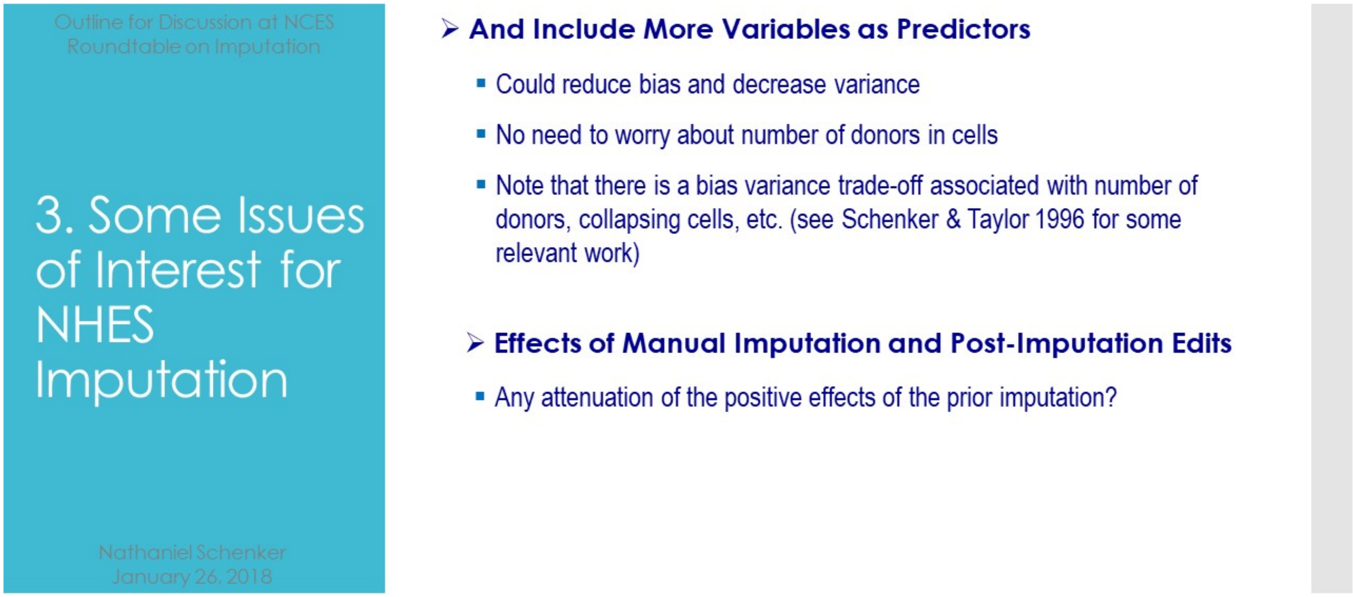
Single Imputation vs Multiple Imputation

- So far, differences in variance estimates not major (note low item nonresponse rates)
- See if there are classes of analyses for which differences are larger

Possible Advantages of Explicit-Model-Based Imputation Over Hot-Deck Imputation

Handles general patterns of missing data better

- Predictors (analogous to "boundary variables") can have missingness
- Note that "random imputation" (used for "boundary variables") probably ok for marginal distributions, but may attenuate multivariate analyses



Outline for Discussion at NCES Roundtable on Imputation

3. Some Issues of Interest for NHES Imputation

Nathaniel Schenker
January 26, 2018

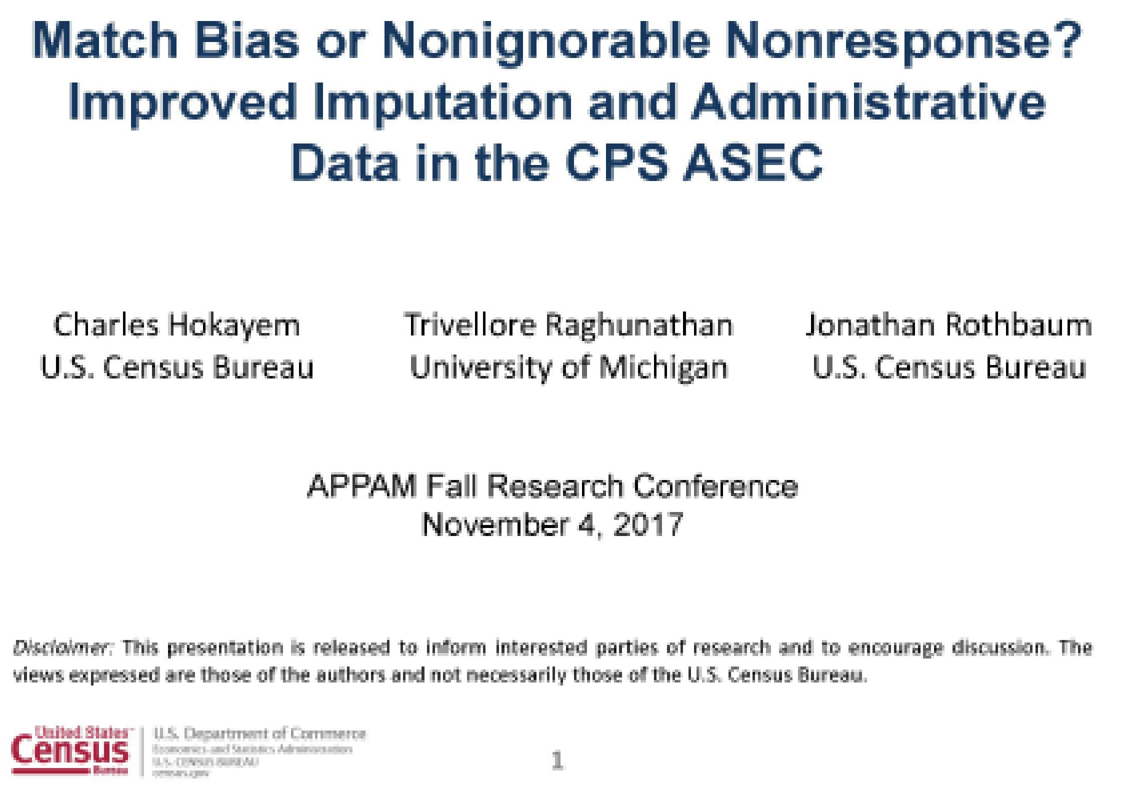
And Include More Variables as Predictors

- Could reduce bias and decrease variance
- No need to worry about number of donors in cells
- Note that there is a bias variance trade-off associated with number of donors, collapsing cells, etc. (see Schenker & Taylor 1996 for some relevant work)

Effects of Manual Imputation and Post-Imputation Edits

- Any attenuation of the positive effects of the prior imputation?

Match Bias or Nonignorable Nonresponse? Improved Imputation & Administrative Data in the CPS ASEC



Match Bias or Nonignorable Nonresponse? Improved Imputation and Administrative Data in the CPS ASEC

Charles Hokayem
U.S. Census Bureau

Trivellore Raghunathan
University of Michigan

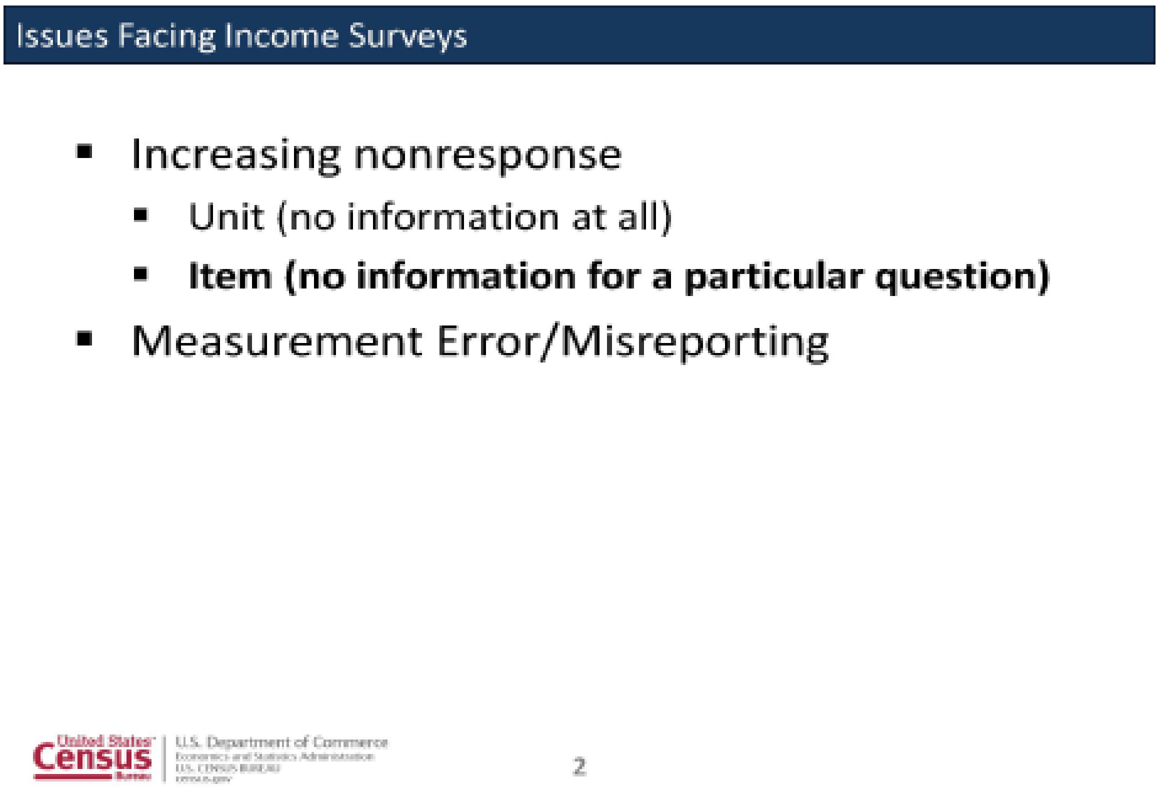
Jonathan Rothbaum
U.S. Census Bureau

APPAM Fall Research Conference
November 4, 2017

Disclaimer: This presentation is released to inform interested parties of research and to encourage discussion. The views expressed are those of the authors and not necessarily those of the U.S. Census Bureau.

United States Census Bureau | U.S. Department of Commerce
Economics and Statistics Administration
U.S. CENSUS BUREAU
census.gov

1

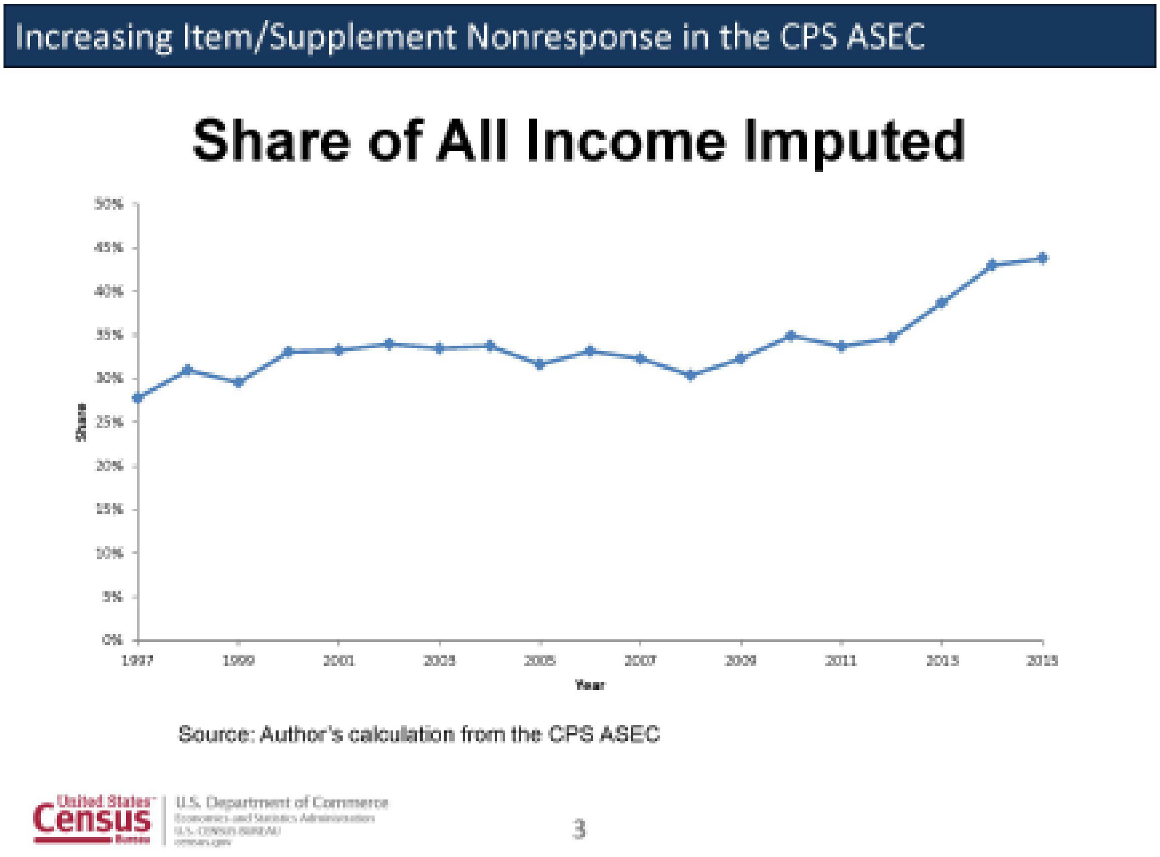


Issues Facing Income Surveys

- Increasing nonresponse
  - Unit (no information at all)
  - **Item (no information for a particular question)**
- Measurement Error/Misreporting

United States Census Bureau | U.S. Department of Commerce
Economics and Statistics Administration
U.S. CENSUS BUREAU
census.gov

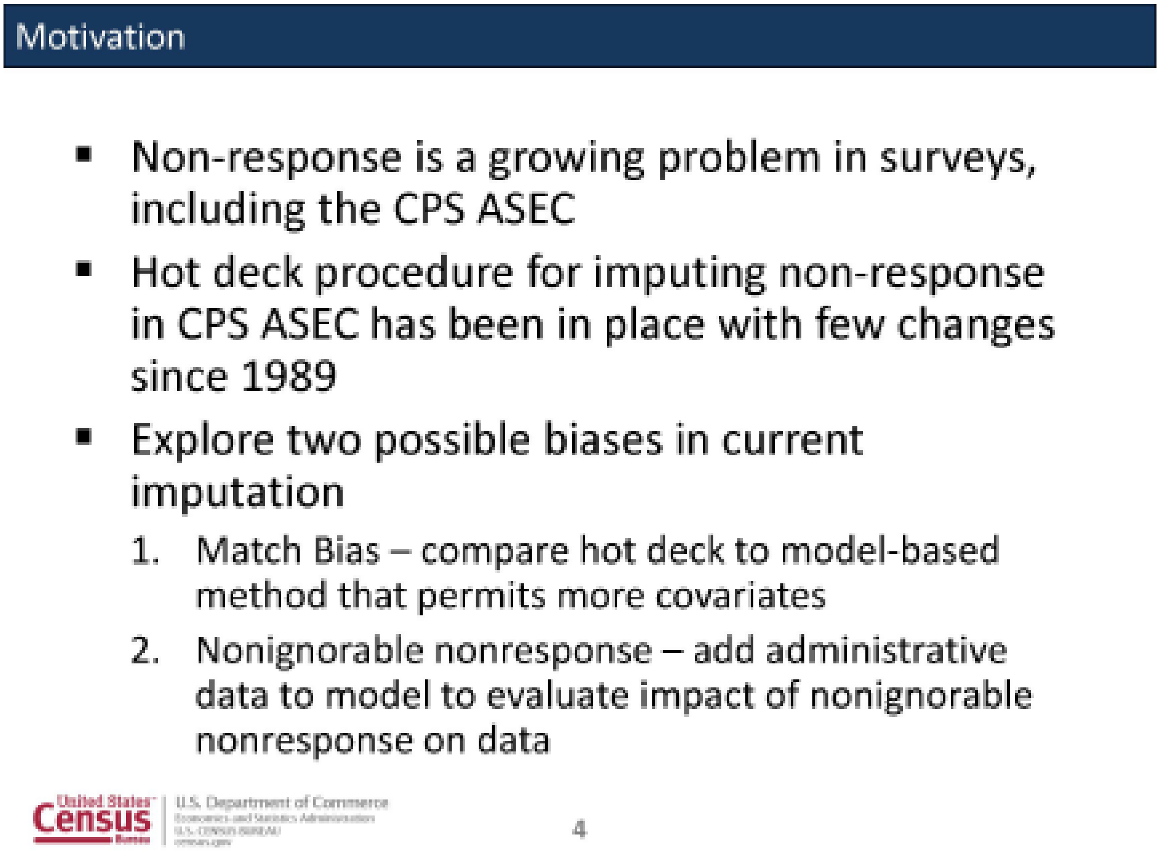
2



Increasing Item/Supplement Nonresponse in the CPS ASEC

Share of All Income Imputed

Source: Author's calculation from the CPS ASEC



Motivation

- Non-response is a growing problem in surveys, including the CPS ASEC
- Hot deck procedure for imputing non-response in CPS ASEC has been in place with few changes since 1989
- Explore two possible biases in current imputation
  1. Match Bias – compare hot deck to model-based method that permits more covariates
  2. Nonignorable nonresponse – add administrative data to model to evaluate impact of nonignorable nonresponse on data

Missing at Random and Match Bias

- Missing at Random – assumed by nearly all imputation models
  - Given Observables O, Unobservables U, and R as response indicator

$$p(R = 1|O, U) = p(R = 1|O)$$

  - For a given statistic Q:

$$E(\hat{Q}|O, U) = E(\hat{Q}|O) = Q$$

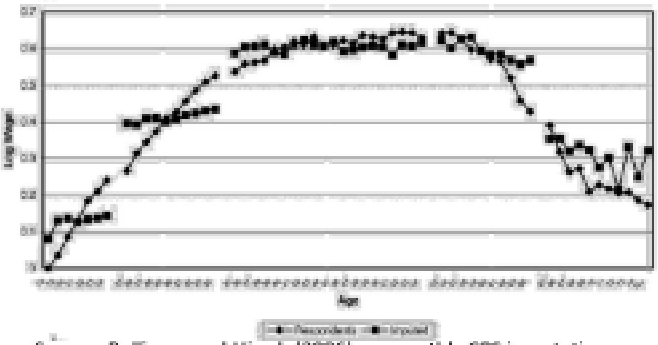
- Match bias – only a subset of variables are in the model (M) and:

$$E(\hat{Q}|O) \neq E(\hat{Q}|M)$$

  - Exclusion of $O_{\setminus \mathrm{M}}$ biases results ($O = \{M, O_{\setminus M}\}$

Match Bias - Examples

- Union status (Hirsch and Schumaker, 2004) – not in CPS imputation model
  - Estimates of wage differences between union/non-union worker attenuated by imputation model's assumption that there is no relationship conditional on M
- Earnings and Experience (Bollinger and Hirsch, 2006) – in CPS model, but grouped
  - Attenuates estimates of returns to experience



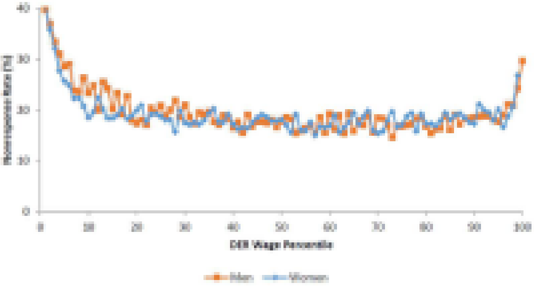
Source: Bollinger and Hirsch (2006) on monthly CPS imputation.

Nonignorable Nonresponse

- Data not missing at random

$$E(\hat{Q}|O,U) \neq E(\hat{Q}|O)$$

  - Exclusion of U biases results
- Example - Trouble in the Tails (Bollinger et al., 2015)
  - Nonresponse is a function of the missing variable, earnings



Source: Bollinger et al. (2015) from CPS ASEC linked to W2 records

Existing Approach: Hot Deck

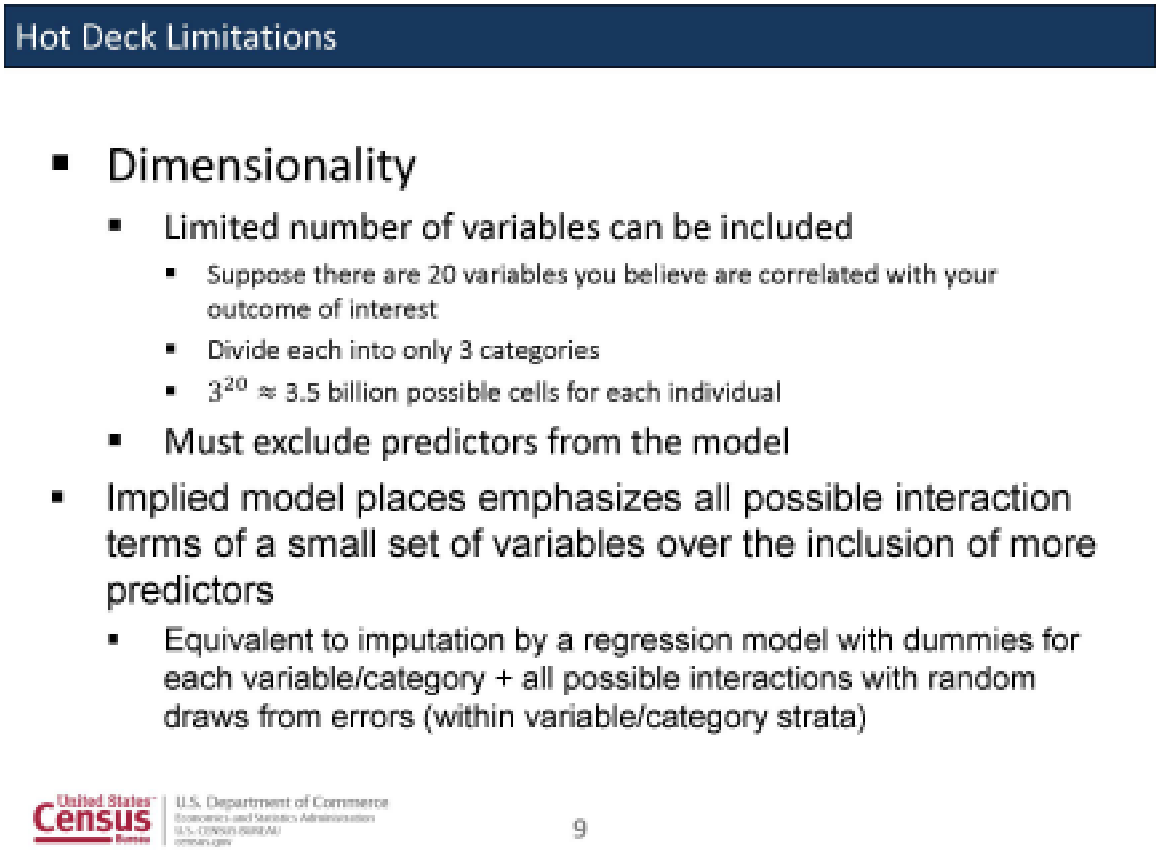
- Match non-respondents to "similar" respondents along a set of characteristics in the model
- Donate response as imputation from respondent to non-respondent

- Example: 2 variables, 2 categories each – 4 cells
  1. Race: White/non-White
  2. Gender: male/female

  Two non-respondents (A and B)

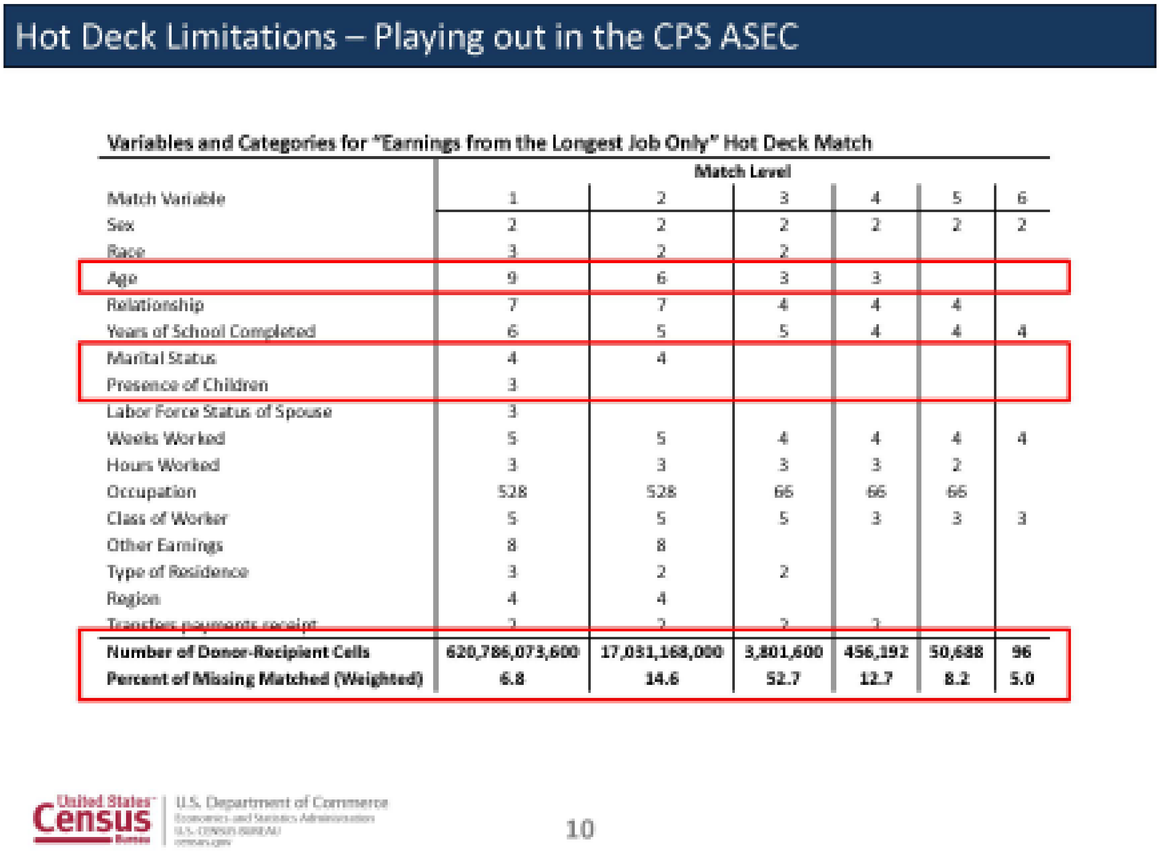
  Person A: white, female – randomly select a white, female respondent and use her response as the imputed value

  Person B: non-white, male – randomly select a non-White, male respondent and use his response as the imputed value



Hot Deck Limitations

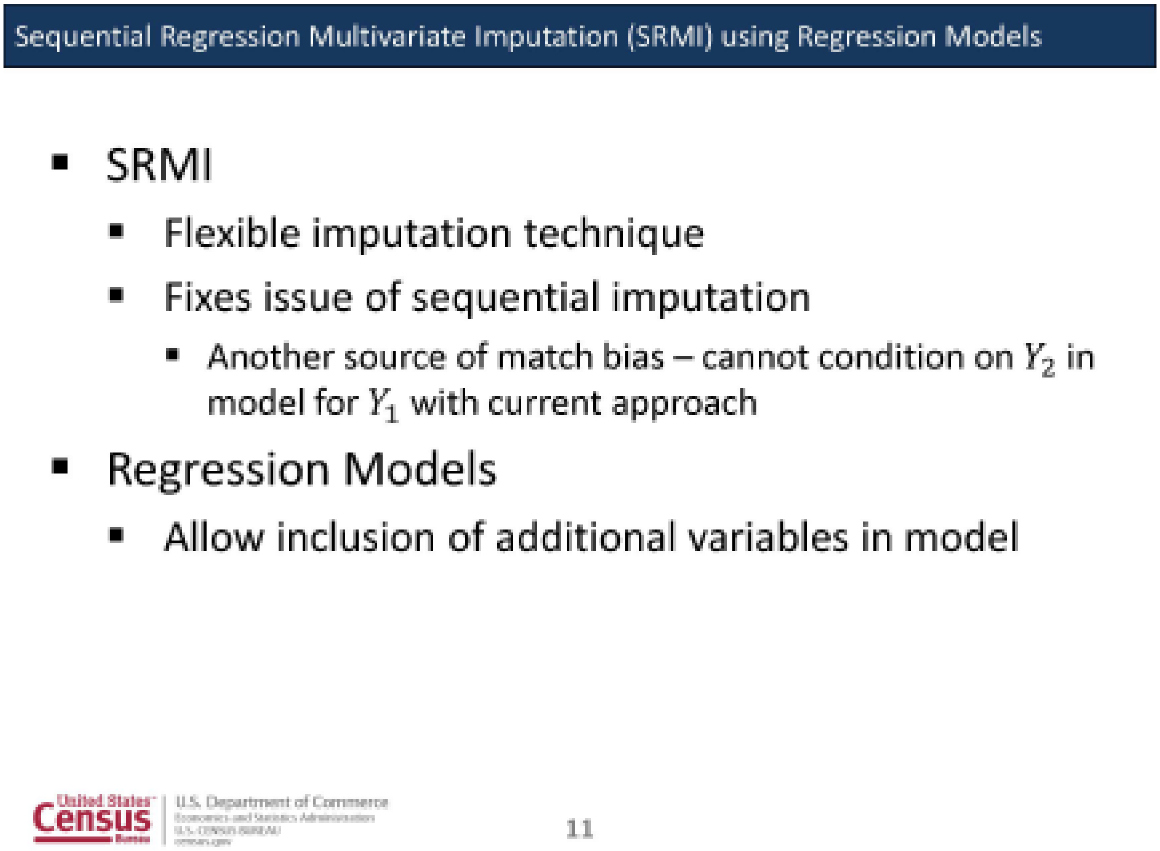
- Dimensionality
  - Limited number of variables can be included
    - Suppose there are 20 variables you believe are correlated with your outcome of interest
    - Divide each into only 3 categories
    - $3^{20} \approx$ 3.5 billion possible cells for each individual
  - Must exclude predictors from the model
- Implied model places emphasizes all possible interaction terms of a small set of variables over the inclusion of more predictors
  - Equivalent to imputation by a regression model with dummies for each variable/category + all possible interactions with random draws from errors (within variable/category strata)



Hot Deck Limitations – Playing out in the CPS ASEC

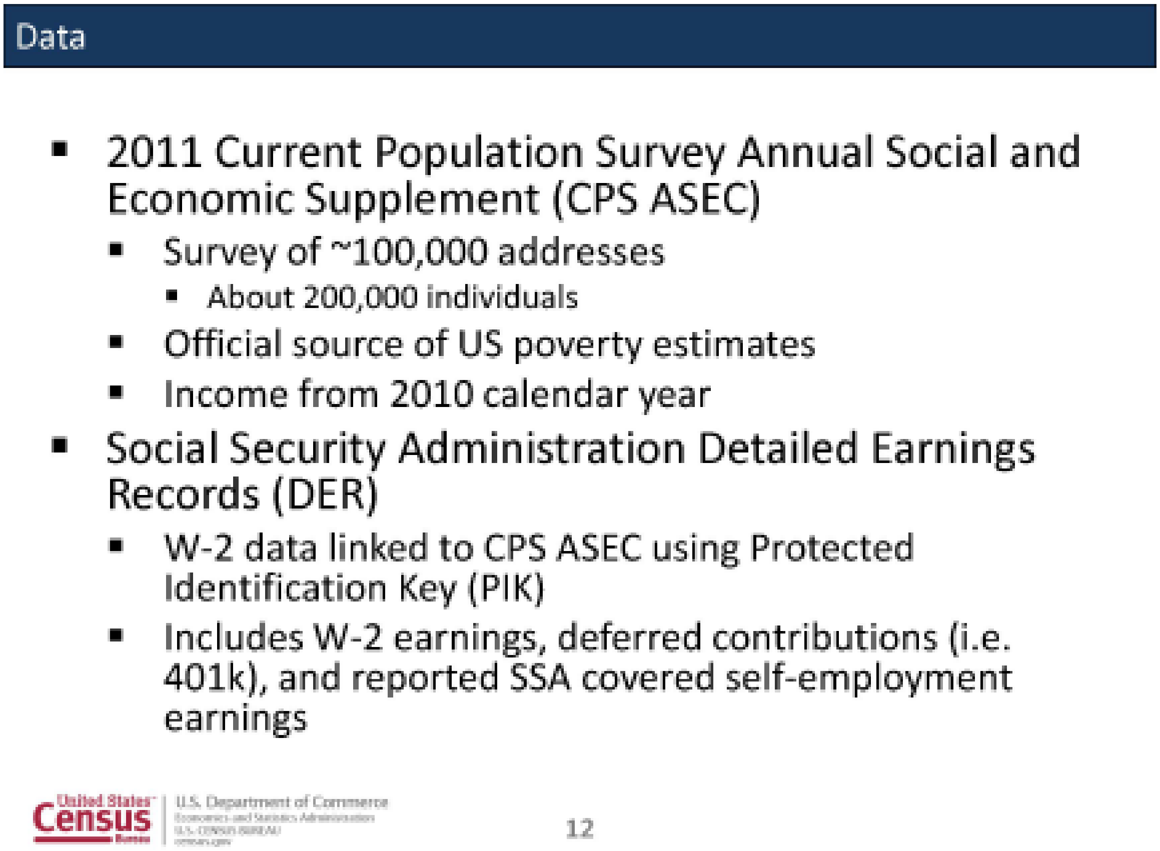
Variables and Categories for "Earnings from the Longest Job Only" Hot Deck Match

	Match Level					
Match Variable	1	2	3	4	5	6
Sex	2	2	2	2	2	2
Race	3	2	2			
Age	9	6	3	3		
Relationship	7	7	4	4	4	
Years of School Completed	6	5	5	4	4	4
Marital Status	4	4				
Presence of Children	3					
Labor Force Status of Spouse	3					
Weeks Worked	5	5	4	4	4	4
Hours Worked	3	3	3	3	2	
Occupation	528	528	66	66	66	
Class of Worker	5	5	5	3	3	3
Other Earnings	8	8				
Type of Residence	3	2	2			
Region	4	4				
Transfers payments receipt	2	2	2	2		
Number of Donor-Recipient Cells	**620,786,073,600**	**17,031,168,000**	**3,801,600**	**456,192**	**50,688**	**96**
Percent of Missing Matched (Weighted)	**6.8**	**14.6**	**52.7**	**12.7**	**8.2**	**5.0**



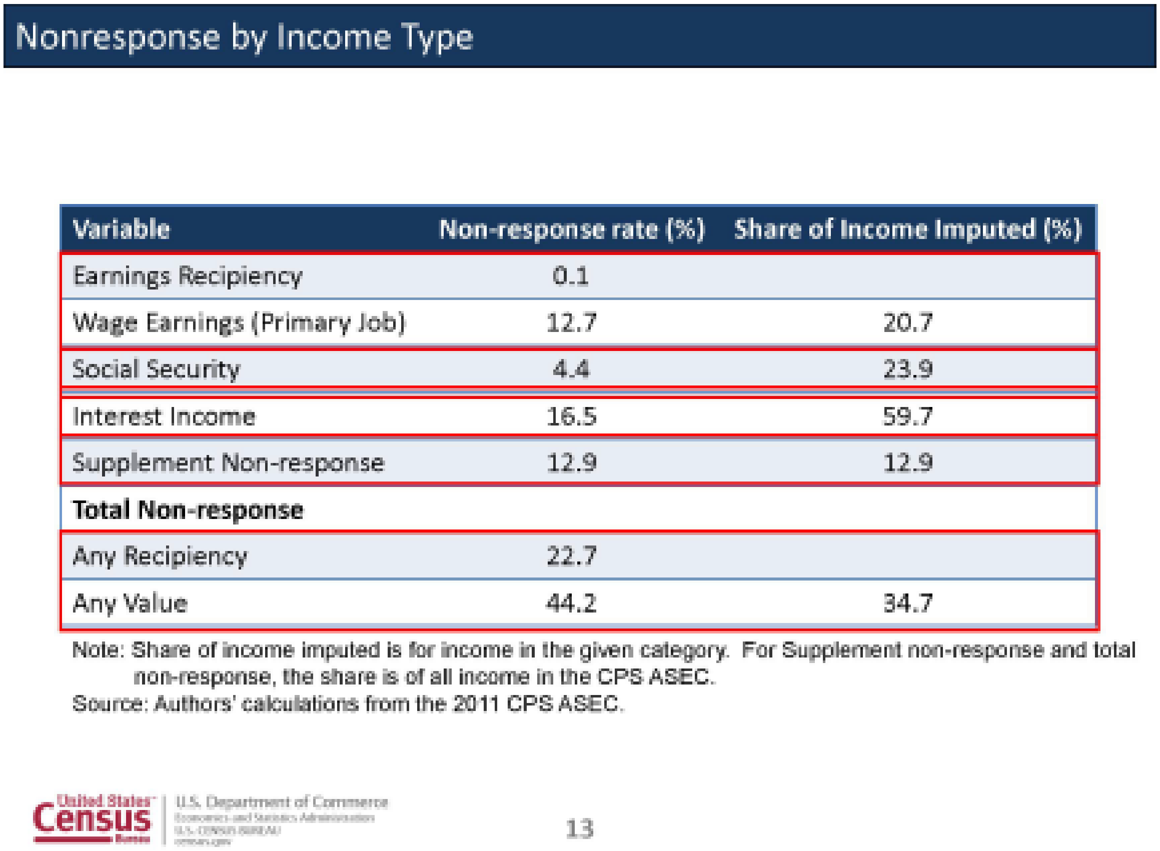
Sequential Regression Multivariate Imputation (SRMI) using Regression Models

- SRMI
  - Flexible imputation technique
  - Fixes issue of sequential imputation
    - Another source of match bias – cannot condition on Y_2 in model for Y_1 with current approach
- Regression Models
  - Allow inclusion of additional variables in model



Data

- 2011 Current Population Survey Annual Social and Economic Supplement (CPS ASEC)
  - Survey of ~100,000 addresses
    - About 200,000 individuals
  - Official source of US poverty estimates
  - Income from 2010 calendar year
- Social Security Administration Detailed Earnings Records (DER)
  - W-2 data linked to CPS ASEC using Protected Identification Key (PIK)
  - Includes W-2 earnings, deferred contributions (i.e. 401k), and reported SSA covered self-employment earnings



Nonresponse by Income Type

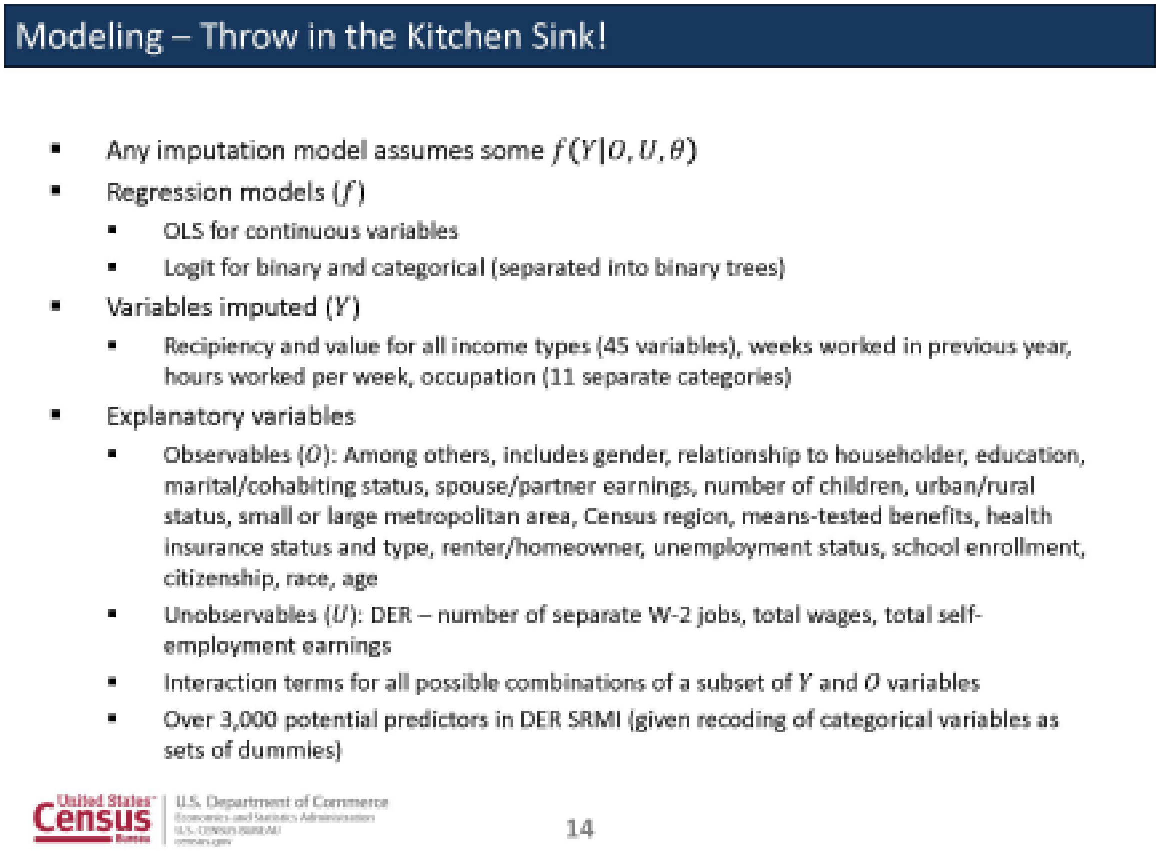
Variable	Non-response rate (%)	Share of Income Imputed (%)
Earnings Recipiency	0.1	
Wage Earnings (Primary Job)	12.7	20.7
Social Security	4.4	23.9
Interest Income	16.5	59.7
Supplement Non-response	12.9	12.9
Total Non-response		
Any Recipiency	22.7	
Any Value	44.2	34.7

Note: Share of income imputed is for income in the given category. For Supplement non-response and total non-response, the share is of all income in the CPS ASEC.

Source: Authors' calculations from the 2011 CPS ASEC.

United States Census Bureau | U.S. Department of Commerce

13

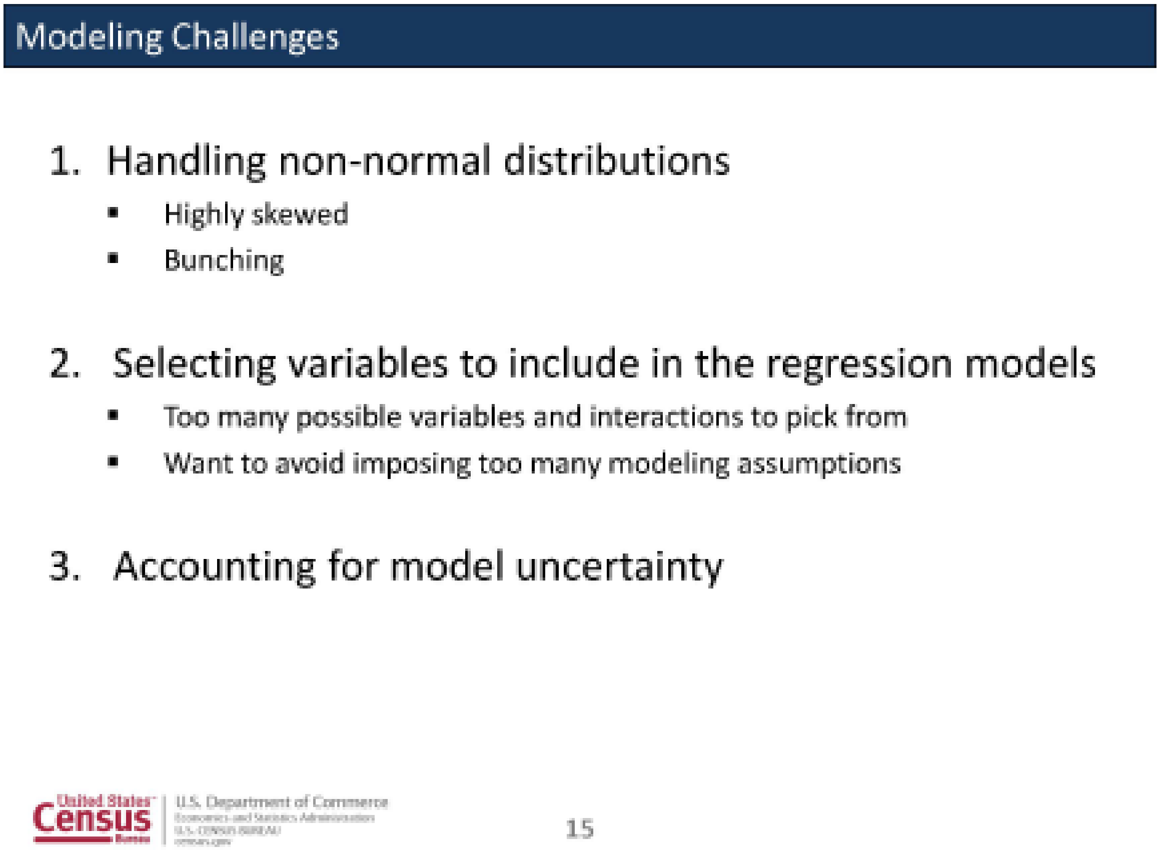


Modeling – Throw in the Kitchen Sink!

- Any imputation model assumes some $f(Y|O,U,\theta)$
- Regression models (f)
  - OLS for continuous variables
  - Logit for binary and categorical (separated into binary trees)
- Variables imputed (Y)
  - Recipiency and value for all income types (45 variables), weeks worked in previous year, hours worked per week, occupation (11 separate categories)
- Explanatory variables
  - Observables (O): Among others, includes gender, relationship to householder, education, marital/cohabiting status, spouse/partner earnings, number of children, urban/rural status, small or large metropolitan area, Census region, means-tested benefits, health insurance status and type, renter/homeowner, unemployment status, school enrollment, citizenship, race, age
  - Unobservables (U): DER – number of separate W-2 jobs, total wages, total self-employment earnings
  - Interaction terms for all possible combinations of a subset of Y and O variables
  - Over 3,000 potential predictors in DER SRMI (given recoding of categorical variables as sets of dummies)

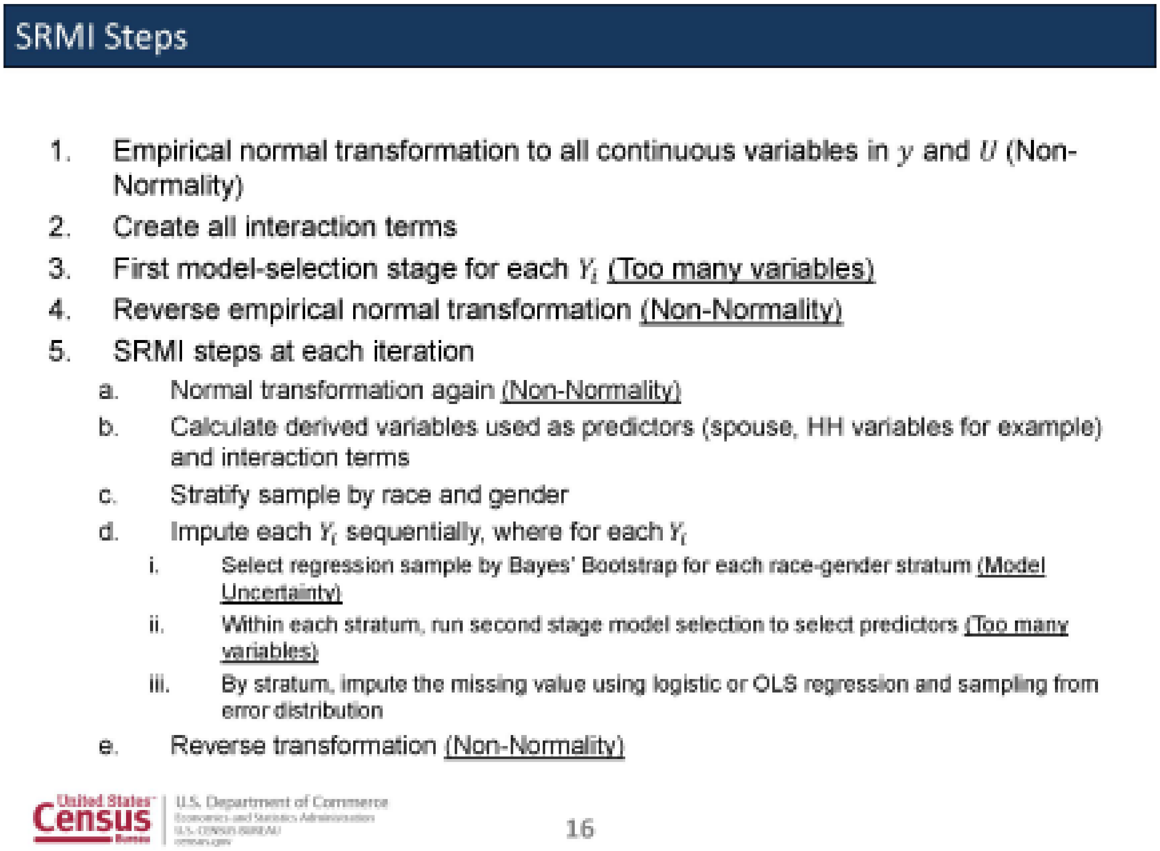
United States Census Bureau | U.S. Department of Commerce

14



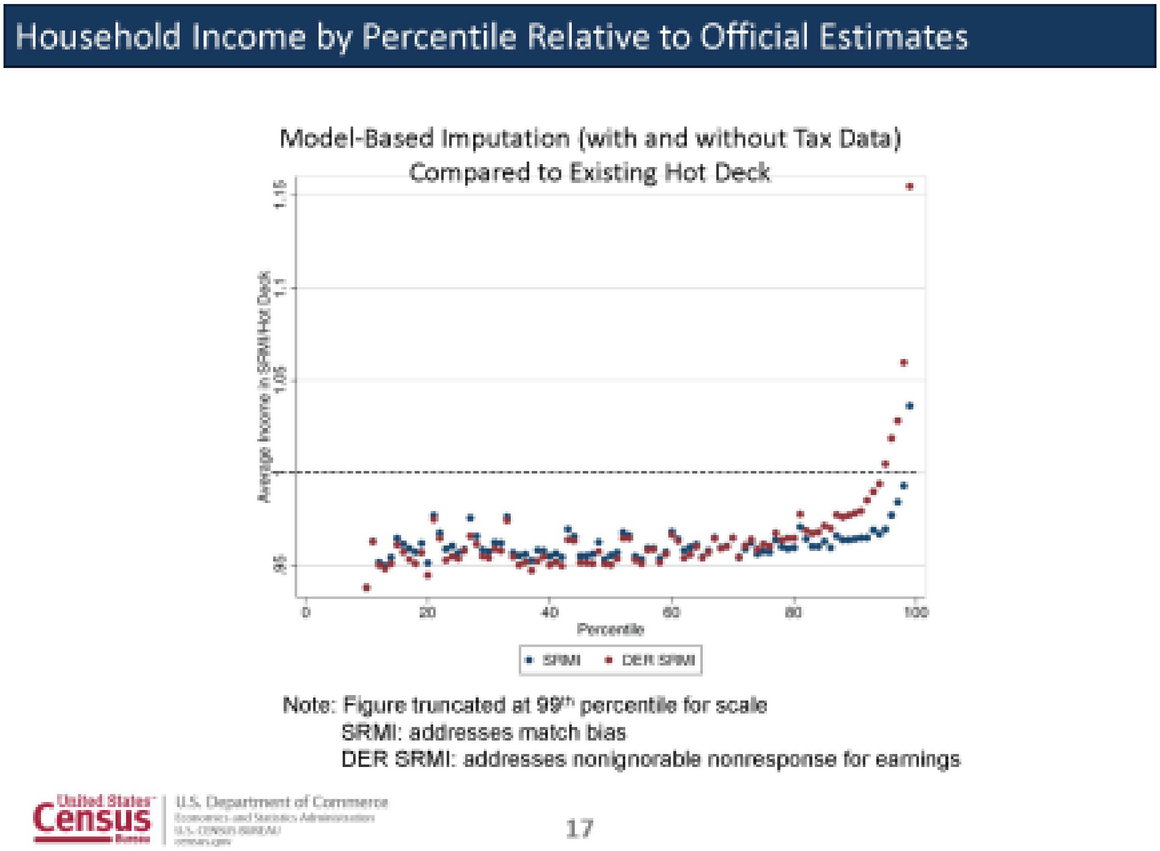
Modeling Challenges

1. Handling non-normal distributions
   - Highly skewed
   - Bunching
2. Selecting variables to include in the regression models
   - Too many possible variables and interactions to pick from
   - Want to avoid imposing too many modeling assumptions
3. Accounting for model uncertainty



SRMI Steps

1. Empirical normal transformation to all continuous variables in y and U (Non-Normality)
2. Create all interaction terms
3. First model-selection stage for each Y_i (Too many variables)
4. Reverse empirical normal transformation (Non-Normality)
5. SRMI steps at each iteration
   a. Normal transformation again (Non-Normality)
   b. Calculate derived variables used as predictors (spouse, HH variables for example) and interaction terms
   c. Stratify sample by race and gender
   d. Impute each Y_i sequentially, where for each Y_i
      i. Select regression sample by Bayes' Bootstrap for each race-gender stratum (Model Uncertainty)
      ii. Within each stratum, run second stage model selection to select predictors (Too many variables)
      iii. By stratum, impute the missing value using logistic or OLS regression and sampling from error distribution
   e. Reverse transformation (Non-Normality)

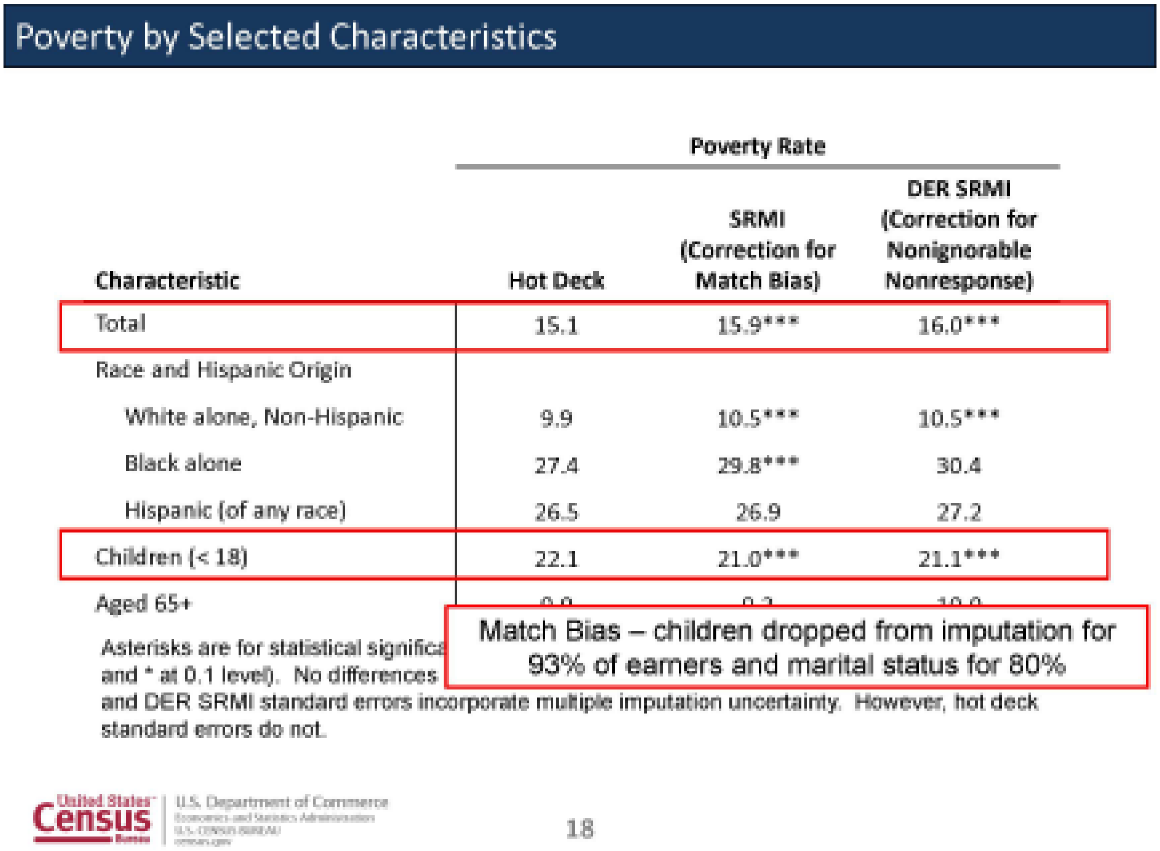


Household Income by Percentile Relative to Official Estimates

Note: Figure truncated at 99th percentile for scale
SRMI: addresses match bias
DER SRMI: addresses nonignorable nonresponse for earnings

United States Census Bureau | U.S. Department of Commerce
Economics and Statistics Administration
U.S. CENSUS BUREAU
census.gov

17



Poverty by Selected Characteristics

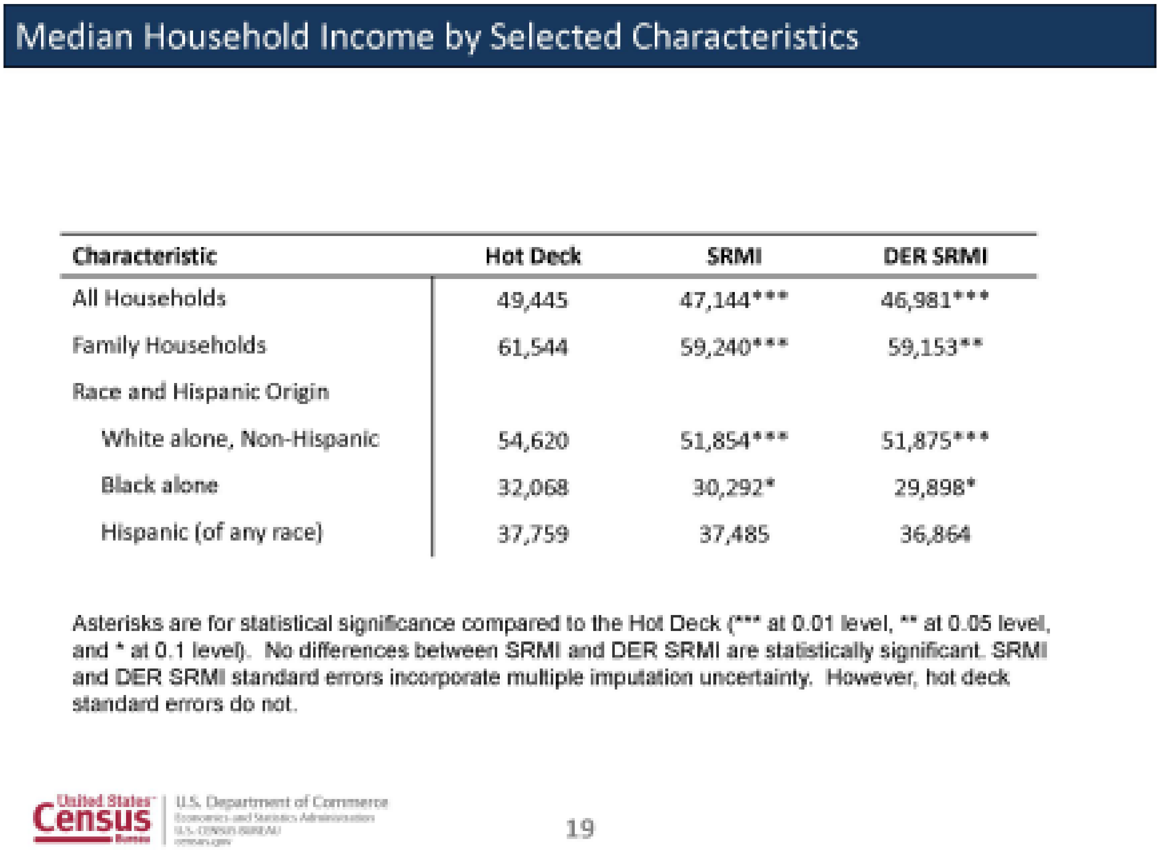
	Poverty Rate		
Characteristic	Hot Deck	SRMI (Correction for Match Bias)	DER SRMI (Correction for Nonignorable Nonresponse)
Total	15.1	15.9***	16.0***
Race and Hispanic Origin			
White alone, Non-Hispanic	9.9	10.5***	10.5***
Black alone	27.4	29.8***	30.4
Hispanic (of any race)	26.5	26.9	27.2
Children (< 18)	22.1	21.0***	21.1***
Aged 65+	[illegible]	[illegible]	[illegible]

Asterisks are for statistical significa
and * at 0.1 level). No differences
and DER SRMI standard errors incorporate multiple imputation uncertainty. However, hot deck standard errors do not.

Match Bias – children dropped from imputation for 93% of earners and marital status for 80%

United States Census Bureau | U.S. Department of Commerce
Economics and Statistics Administration
U.S. CENSUS BUREAU
census.gov

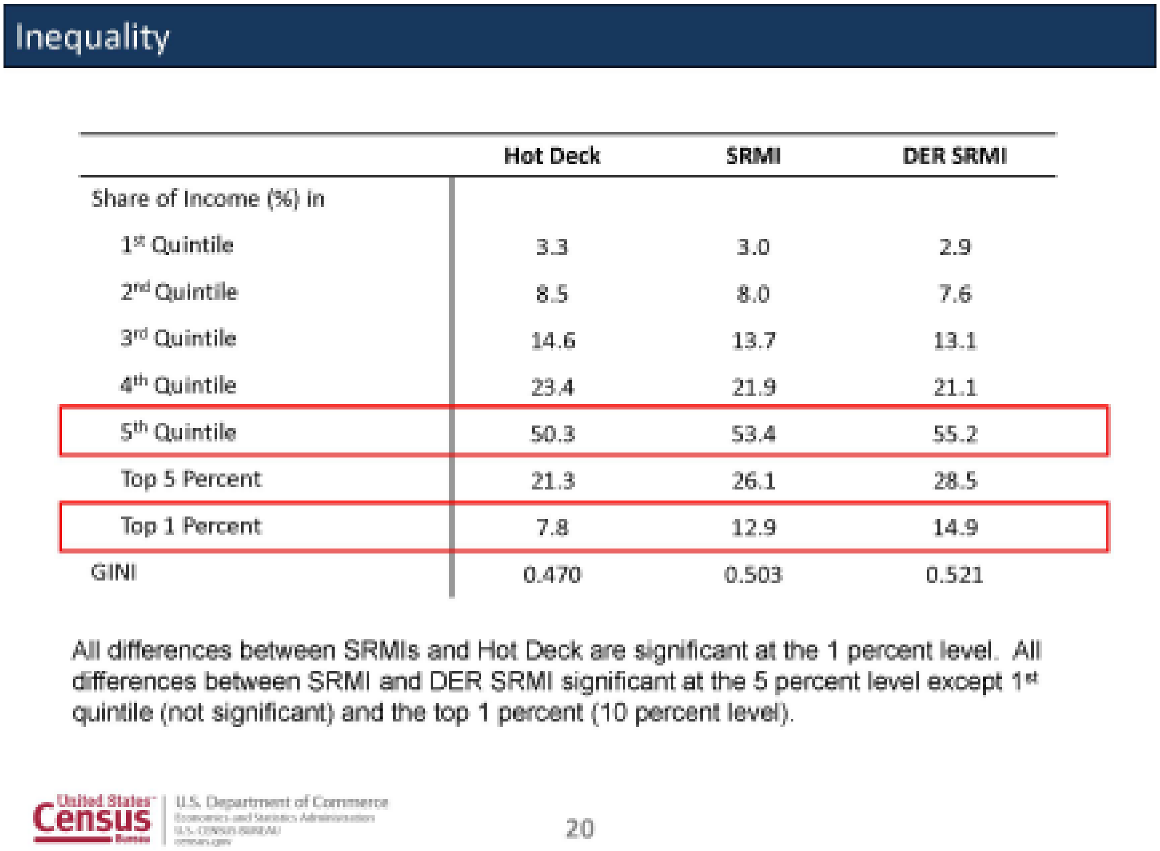
18



Median Household Income by Selected Characteristics

Characteristic	Hot Deck	SRMI	DER SRMI
All Households	49,445	47,144***	46,981***
Family Households	61,544	59,240***	59,153**
Race and Hispanic Origin			
White alone, Non-Hispanic	54,620	51,854***	51,875***
Black alone	32,068	30,292*	29,898*
Hispanic (of any race)	37,759	37,485	36,864

Asterisks are for statistical significance compared to the Hot Deck (*** at 0.01 level, ** at 0.05 level, and * at 0.1 level). No differences between SRMI and DER SRMI are statistically significant. SRMI and DER SRMI standard errors incorporate multiple imputation uncertainty. However, hot deck standard errors do not.

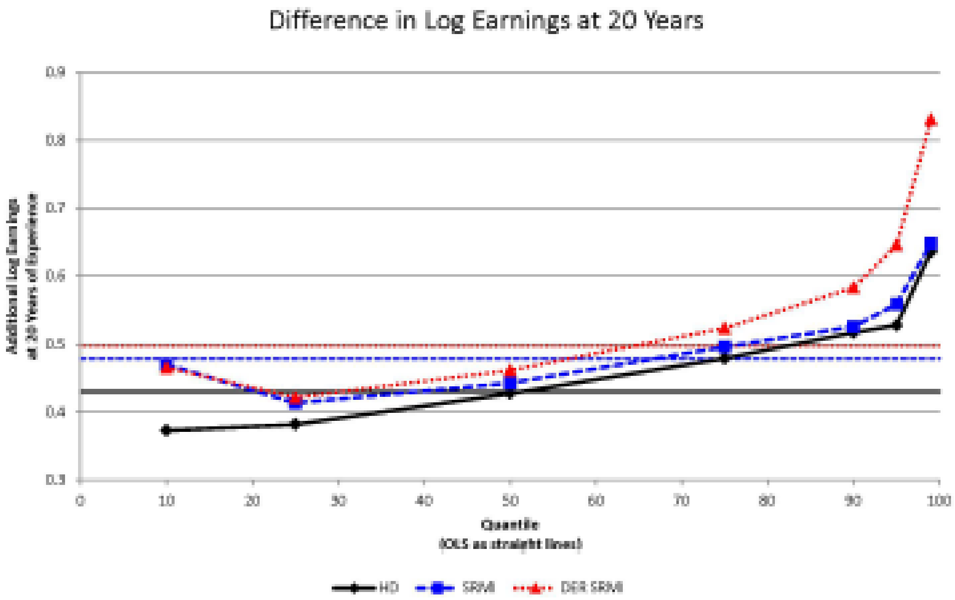


Inequality

	Hot Deck	SRMI	DER SRMI
Share of Income (%) in			
1st Quintile	3.3	3.0	2.9
2nd Quintile	8.5	8.0	7.6
3rd Quintile	14.6	13.7	13.1
4th Quintile	23.4	21.9	21.1
5th Quintile	50.3	53.4	55.2
Top 5 Percent	21.3	26.1	28.5
Top 1 Percent	7.8	12.9	14.9
GINI	0.470	0.503	0.521

All differences between SRMIs and Hot Deck are significant at the 1 percent level. All differences between SRMI and DER SRMI significant at the 5 percent level except 1st quintile (not significant) and the top 1 percent (10 percent level).

Returns to Experience (Mincer Earnings Regression)

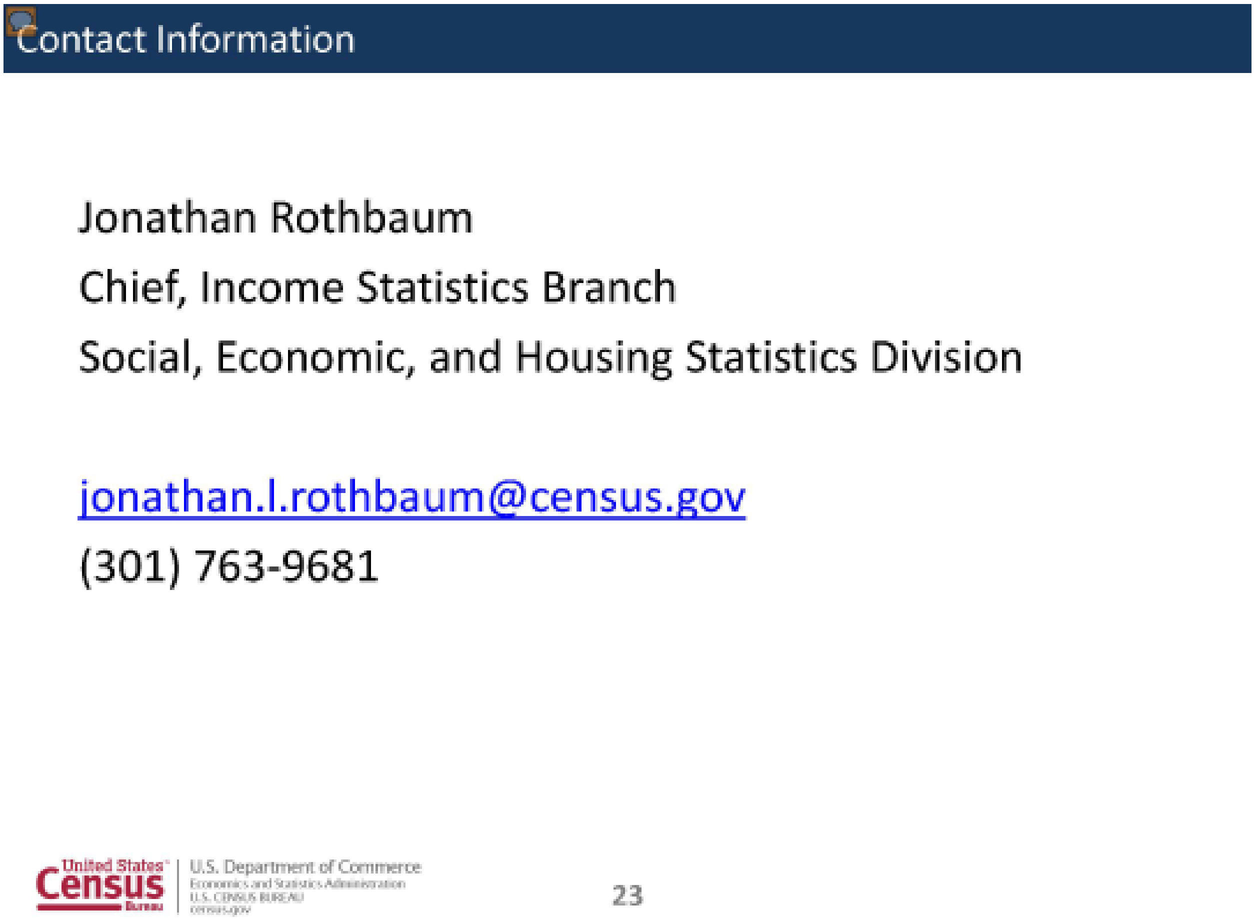


Future Research

1. Add more sources of administrative records
   - 1040 information
   - 1099Rs
   - SSA Records – OASDI and SSI payments
   - State-provided means-tested program benefits

2. Add more years to understand if/how nonresponse bias has changed over time

3. Include more summary information by geography to better capture associations between state and local area characteristics



Contact Information

Jonathan Rothbaum
Chief, Income Statistics Branch
Social, Economic, and Housing Statistics Division

jonathan.l.rothbaum@census.gov
(301) 763-9681

Appendix A: Research Technical Experts

Research Technical Experts

Moon Jung Cho, PhD
Office of Survey Methods Research, US Bureau of Labor Statistics

Jae Kwang Kim, PhD
Professor, Department of Statistics & Center for Survey Statistics & Methodology (CSSM), Iowa State University

Trivellore Raghunathan, PhD
Professor of Biostatistics, Director & Research Professor, Survey Research Center, Institute for Social Research, University of Michigan-Ann Arbor

Nathaniel Schenker, PhD
Consultant

Clifford Spiegelman, PhD
Distinguished Professor, Department of Statistics, Texas A&M University

Yves Thibaudeau, PhD
Principal Researcher, Center for Statistical Research & Methodology, US Census

Recai Yucel, PhD
Chair & Associate Professor of Biostatistics, Department of Epidemiology & Biostatistics, School of Public Health, SUNY-Albany

Panel convened by National Institute of Statistical Sciences

Nell Sedransk, PhD
Director-DC, NISS

Ya Mo, PhD
Research Associate, NISS-DC

Funding Sponsor

National Center for Education Statistics