That's a Long Survey!

Using Split-Questionnaire Design to Reduce Respondent Burden in a State Health Survey

Cameron McPhee | SSRS

Cordelia Horch | SSRS

YuChing Yang | University of California – Los Angeles

Jiangzhou Fu | University of California – Los Angeles

Federal Committee on Statistical Methodology

Acknowledgements/Disclaimer

We thank the **University of California – Los Angeles Center for Health Policy Research** for allowing the use of California Health Interview Survey data for this presentation.

Views expressed in this presentation are those of the presenter and not the study sponsor nor any organization with which the authors are affiliated.

ssrs
redefining research

Background

Surveys are Long…Especially State and Federal Surveys

ssrs redefining research

National Health Interview Survey (Since 1957): 60 MIN

NSFG: 75 MIN

The General Social Survey GSS: 90 MIN

HRS Health and Retirement Study: 2 to 3 HOURS

1. **Length Impacts Burden**
2. **Burden Impacts Cooperation**
3. **Cooperation Impacts Data Quality and Cost**

> 20,000 unique respondent per year

>700 Unique Survey Questions

Average Adult Survey Length in CHIS:

- 2021 was 49 Minutes
- Web Mode Average is 47 Minutes
- Phone Mode Average is 69 Minutes

Non-English Surveys Took Longer

(~58 Minutes)

Individual Section Timings Ranged from Less than 1 Minute to About 6 Minutes

Split Questionnaire Designs (*a.k.a. Matrix-Sampling*)

Methods

- Respondents (rows) and items (columns) are *both* "sampled" from a conceptual complete population data matrix
- Ideally creates missing at random (MAR) or even MCAR data distributions

Benefits

- Recover univariate and multivariate distributions with, often, minimal loss in precision (Raghunathan & Grizzle, 1995), though findings are mixed (Axenfeld, et., al., 2022)
- Reduce measurement error associated with longer questionnaires (Peytchev & Peytcheva, 2017)

Constraints

- Must include good predictors of the split items in core modules (Thomas, et. al., 2006)
- Assignment of items to modules can impact the bias introduced through imputation (Axenfeld, et., al., 2022)

ssrs
redefining research

Split-Questionnaire Design (SQD) and CHIS

Research Questions

Is there a modular (split-questionnaire) structure that could be implemented for the CHIS that…

- Meaningfully **reduces burden** on respondents?
- Preserves the **logical flow** of the survey instrument
- Includes "core" questionnaire items that are **highly correlated** to the items being imputed
- Maintains all **weighting** items in the core questionnaire
- Produces unbiased estimates overall and for key subgroups relative to the full questionnaire design

- **Allows for the construction of a usable, fully-imputed, publicly releasable data file (i.e., does not require multiple imputation)**
- Will not attenuate variance estimates relative to the full questionnaire design

Methods: Form Design

1. **Created Modules:** Divided the CHIS questionnaire into topic clusters (modules) of approximately similar numbers of items
2. **Identified Core Items:** For each topic cluster, identified "core" items
   - Often these were "routing" items that determines the follow-up questions a respondent receives
3. **Built Forms:** Created 28 unique "forms", each included
   - All demographic survey questions
   - All items identified as core items for all modules
   - All non-core items for two modules
4. **Selected Critical Variables:** Chose 21 "key" survey items for simulation and analysis

Modules, Routing Items, and Dependent Items

	1. Health Conditions and Disability	2. Smoking, Alcohol, and Drug Use	3. Health and Sexual Behavior
CORE ITEMS	• General health • Diagnoses: Asthma. Diabetes, etc. • COVID related questions • Tested for colon cancer • Vision/hearing difficulties	• 100+ Cigarettes • E-cig. or vape use • Chewing tobacco/Snuff (30 days) • Marijuana/CBD use • Heroin (12 months) • Alcohol (ever)	• How often eat fruit • Importance of genetics and medical care • # of firearms in home • # of sexual partners (12 months) • Sexual orientation • Ever used PrEP • Ever tested for HIV • Ever received HPV vaccine
DEPENDENT ITEMS	• COVID vaccine receipt	• Person around you smokes/vapes • Days drank 4+ alcoholic drinks	• # sweet beverages (past month) • Importance of environmental & behavioral factors to health • Was offered HIV test

Modules, Routing Items, and Dependent Items

	4. Health Insurance Coverage	5. Health Insurance Detail	6. Health Care
CORE ITEMS	• Medicare/Medi-CAL • Employer Insurance • Private insurance • CHAMPUS/CHAMP-VA, or military insurance • Other government health insurance program • Other health insurance	• HMO • High deductible • Continual insurance (12 months) • Reason for uninsurance • Reached plan limit • Reason not enrolled in Medi-CAL	• Usual place for care • ER use • Hospital stays • # of doctor's visits • Telehealth use • Difficulty understanding physician • Difficulty/delay receiving medication • Pregnancy status/plans • Mammogram • Dental visits • Racial barriers to care • Need for mental health care
DEPENDENT ITEMS	• Monthly cost of health plan	• Prescription drug coverage • Deductible over $2,000 • Previously had health coverage	• Time since last checkup • Telehealth in past 12 months • Dental insurance

Modules, Routing Items, and Dependent Items

	7. Psychological Distress/Mental Health	8. Employment, Housing, & Earnings
CORE ITEMS	• Felt nervous, hopeless, restless, depressed, everything was an effort, worthless (past 30 days) • Experienced hazardous climate event • Intimate partner violence • Live with anyone depressed or mentally ill • Able to talk about feelings growing up • # times stopped by police (past 3 years) • Suicidal thoughts (self/close friends)	• Work hours • Length at job • Earnings, income, child support, worker's comp, Social Security/pension • Awareness of CA FMLA laws • Taken paid leave for more than 2 weeks • Receiving TANF or CalWORKS • Housing unit type, tenure, length at address • Help neighbors, neighbors get along, neighbors can be trusted • Volunteered in community • Did not apply for services due to self/family immigration status
DEPENDENT ITEMS	• Phone/computer use (per day) • Physical abuse from intimate partner • Importance of providers asking ACEs • 2 non-parent adults involved in childhood • Ever been arrested	• Feelings about current housing situation • Feel safe in neighborhood • Why not vote in most recent election

Form Splits

ssrs redefining research

Form 1
- Demographics
- Routing Items (all)
- Module 1
- Module 2

Form 2
- Demographics
- Routing Items (all)
- Module 1
- Module 3

Form 3
- Demographics
- Routing Items (all)
- Module 1
- Module 4

Form 4
- Demographics
- Routing Items (all)
- Module 1
- Module 5

Form 5
- Demographics
- Routing Items (all)
- Module 1
- Module 6

Form 17
- Demographics
- Routing Items (all)
- Module 3
- Module 7

Form 18
- Demographics
- Routing Items (all)
- Module 3
- Module 8

Form 19
- Demographics
- Routing Items (all)
- Module 4
- Module 5

Form 20
- Demographics
- Routing Items (all)
- Module 4
- Module 6

Form 21
- Demographics
- Routing Items (all)
- Module 4
- Module 7

Form 24
- Demographics
- Routing Items (all)
- Module 5
- Module 7

Form 25
- Demographics
- Routing Items (all)
- Module 5
- Module 8

Form 26
- Demographics
- Routing Items (all)
- Module 6
- Module 7

Form 27
- Demographics
- Routing Items (all)
- Module 6
- Module 8

Form 28
- Demographics
- Routing Items (all)
- Module 7
- Module 8

Federal Committee on Statistical Methodology

Methods: Simulation and Analysis

Created 50 Replicates of the CHIS:2021 Adult Data File with all Demographic, Routing, and Outcome Variables

- Randomly assigned cases to 1 of 28 forms for each replicate
- Set to missing responses for all questions not included on the assigned form

<u>Phase 1</u>

1. For each dependent variable, use CART model (R ctree) to predict variable value using demographics and all routing items from all sections *other* than the one for that item
   – Export final "node" from the CART model, and assign to all cases in the full file
2. Run sequential hotdeck imputation on missing cases of each dependent variable using section-specific routing items and final CART node to form imputation classes
3. Repeat imputation (allowing imputed cases to donate values) for remaining missing data, under the same model

Methods: Simulation and Analysis (continued)

Phase 2*:

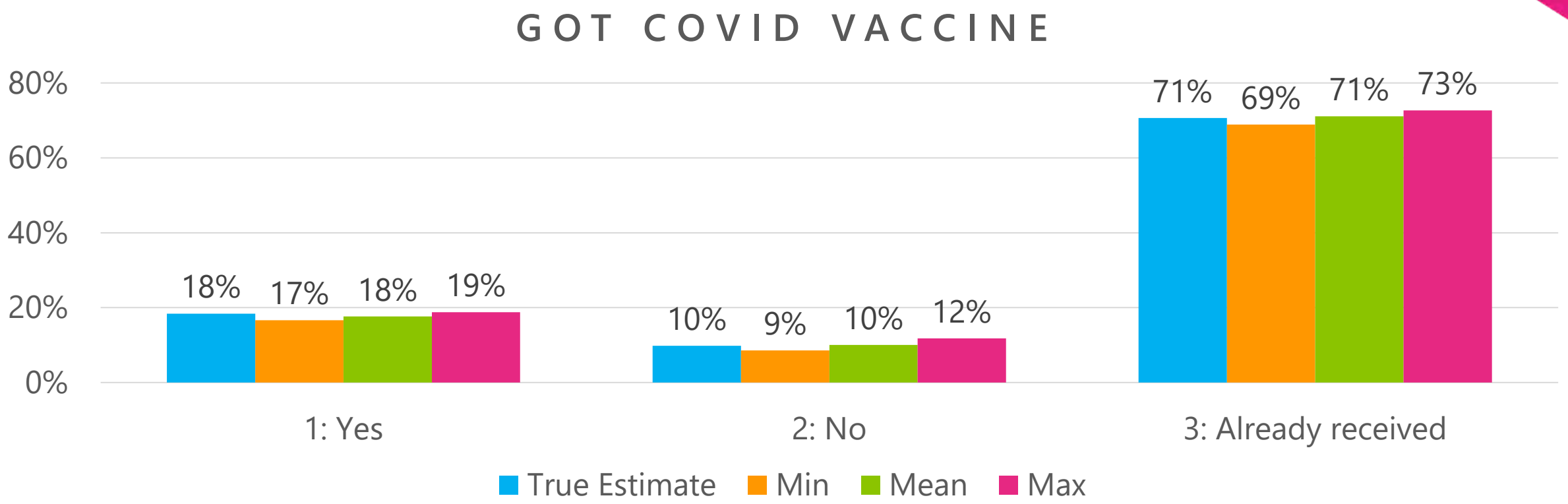
- Once all dependent variables have been imputed once
  1. Blank imputed values (one variable at a time) from all cases originally missing
  2. Rerun CART *and* hotdeck using all variables (including other imputed variables)
  3. Repeat imputation process 5 times (see Marker, Judkins, and Winglee, 2002).

*Phase 2 was not performed for the analyses presented here due to time constraints, but is expected to improve outcomes.

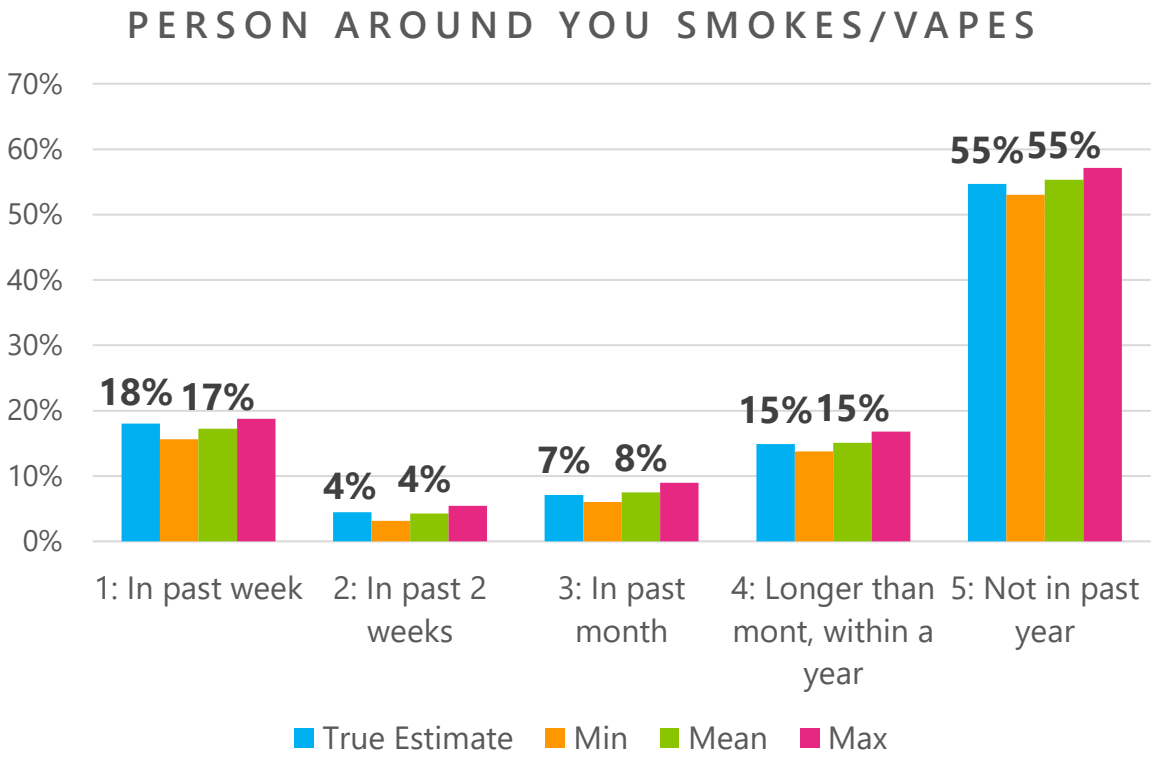
Results: Form Length

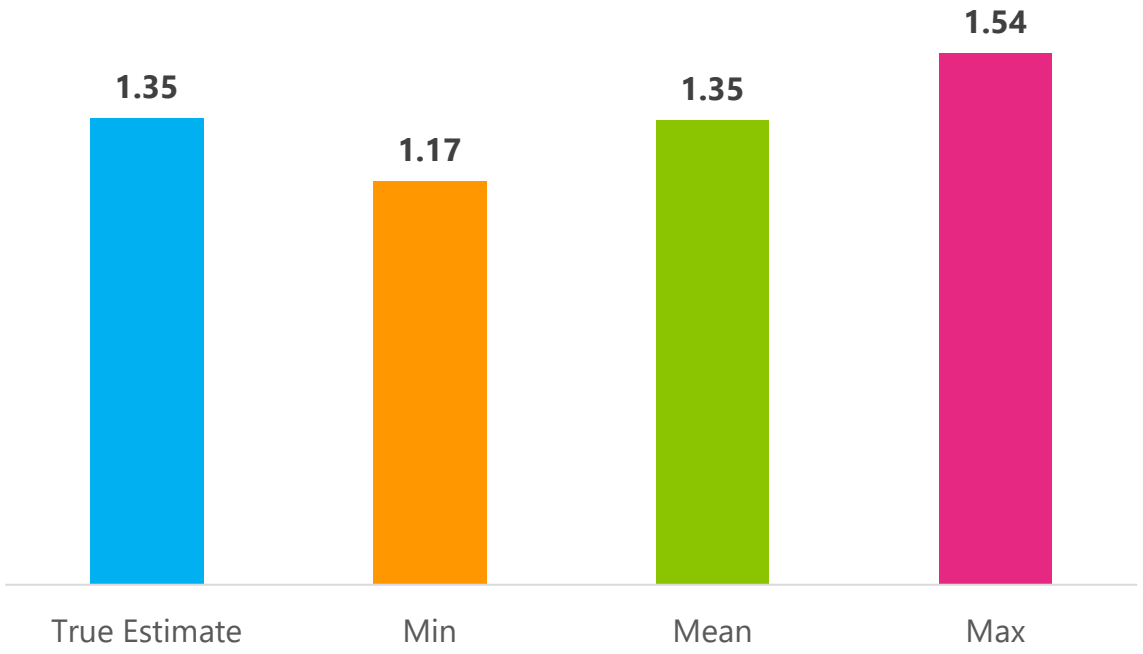
- All forms theoretically would include approx. 80 demographic items & 95 section routing items
- Modules range from 25-89 items
- Form length ranges from 230-345 items
- Respondents complete approximately 10-12 questions per minute (on average)
- New design would yield a survey approximately 25 minutes on average (20-30 minutes depending on the form)
- **Near 50% reduction in survey length!**

Health Conditions and Disability Full Sample Estimates

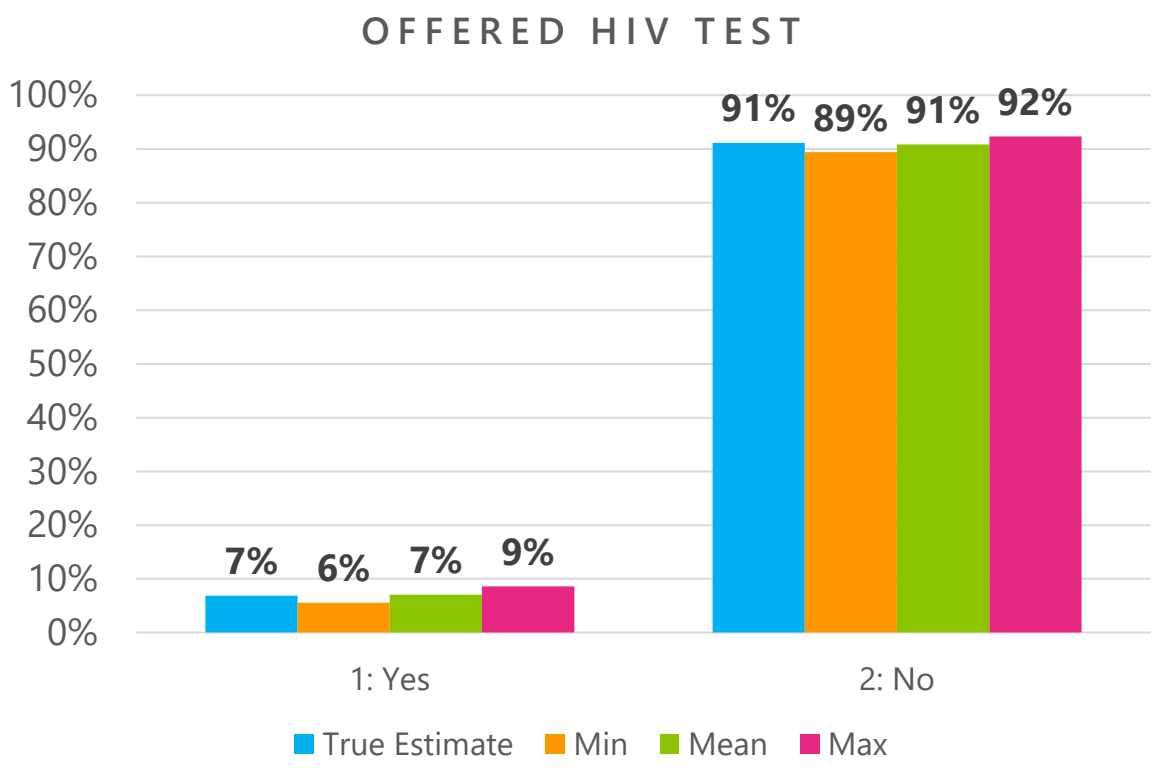


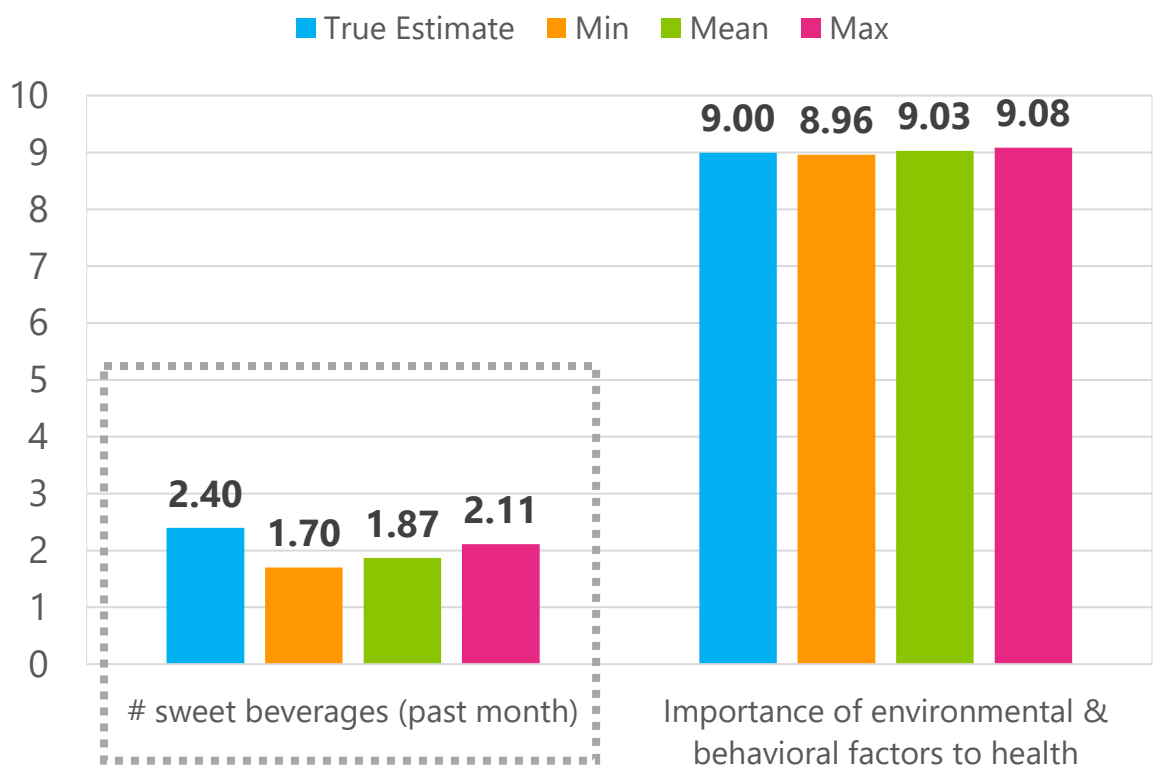
Smoking, Alcohol, and Drug Use Full Sample Estimates





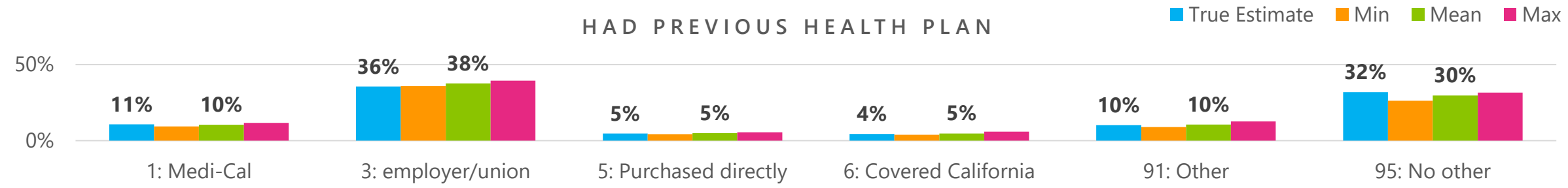
Health and Sexual Behavior
Full Sample Estimates

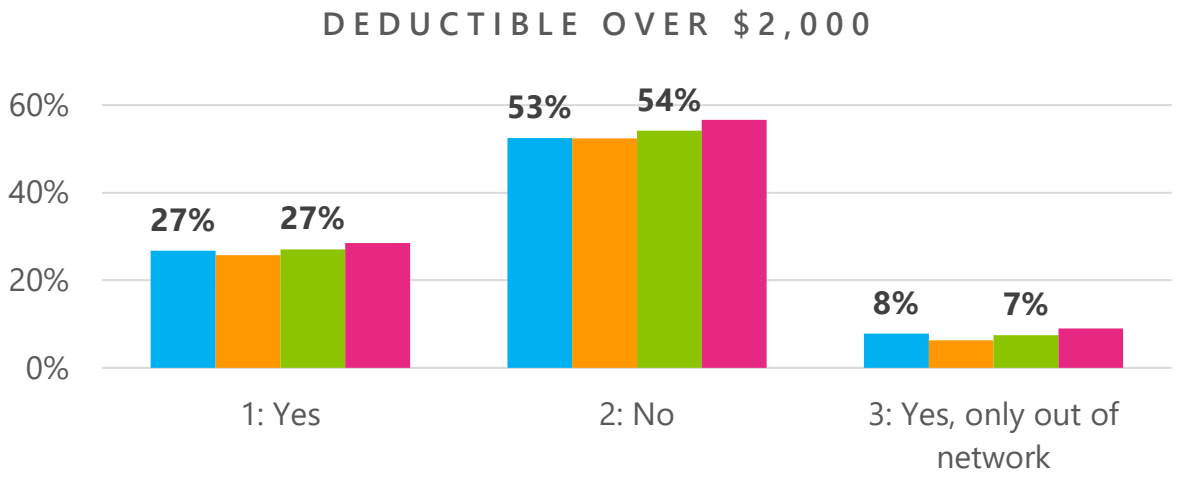


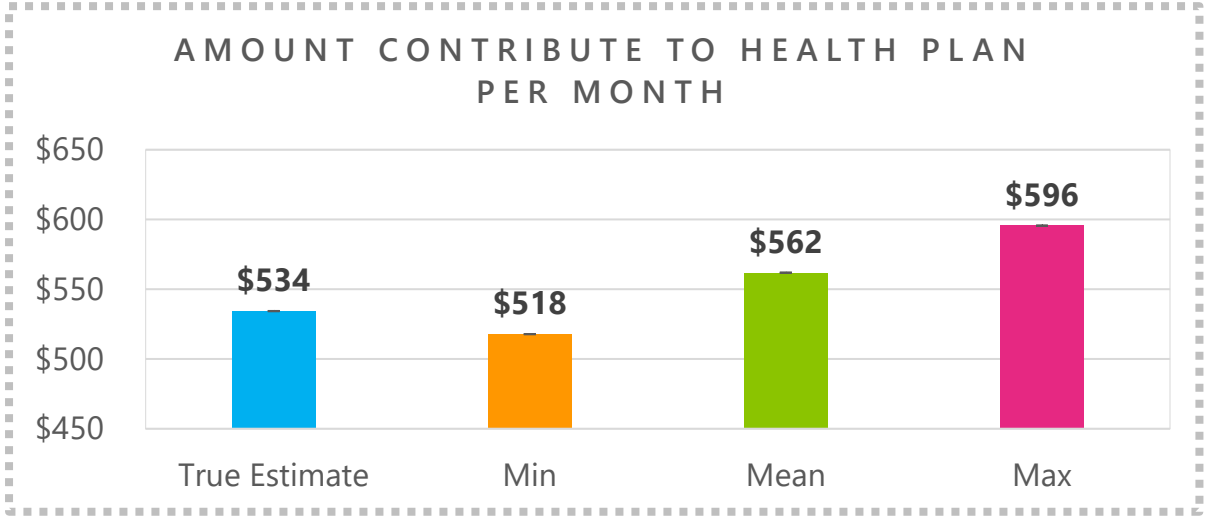


Federal Committee on Statistical Methodology

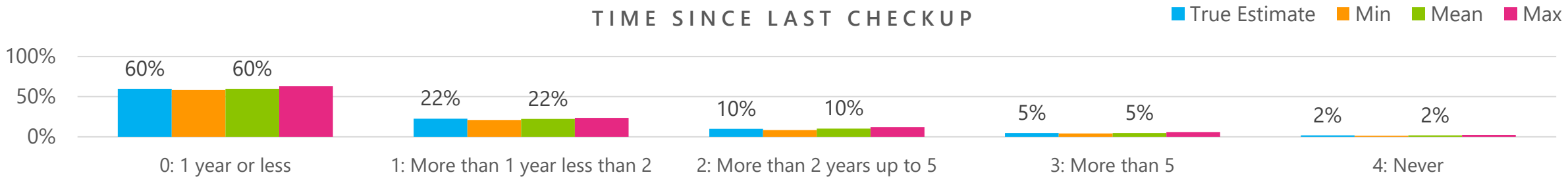
Health Insurance Coverage & Detail
Full Sample Estimates

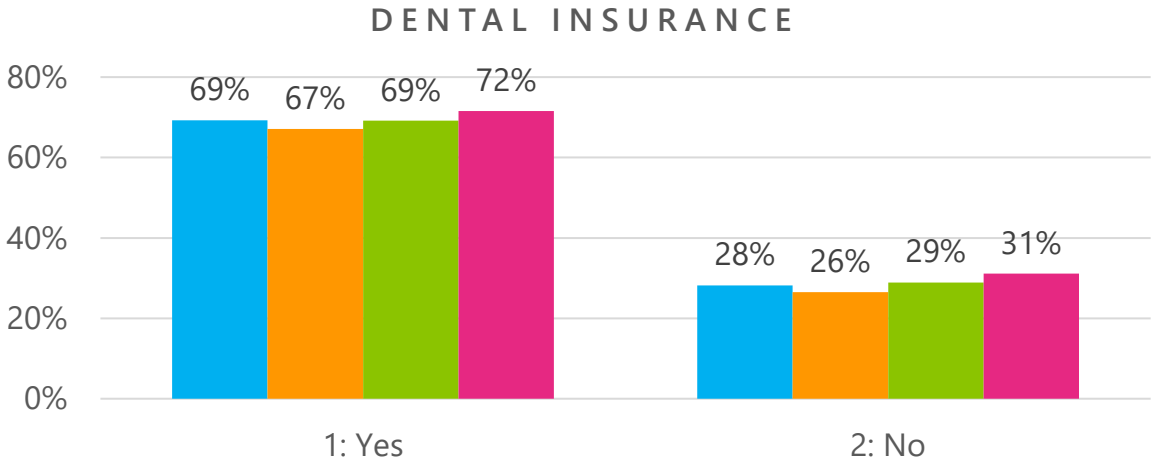


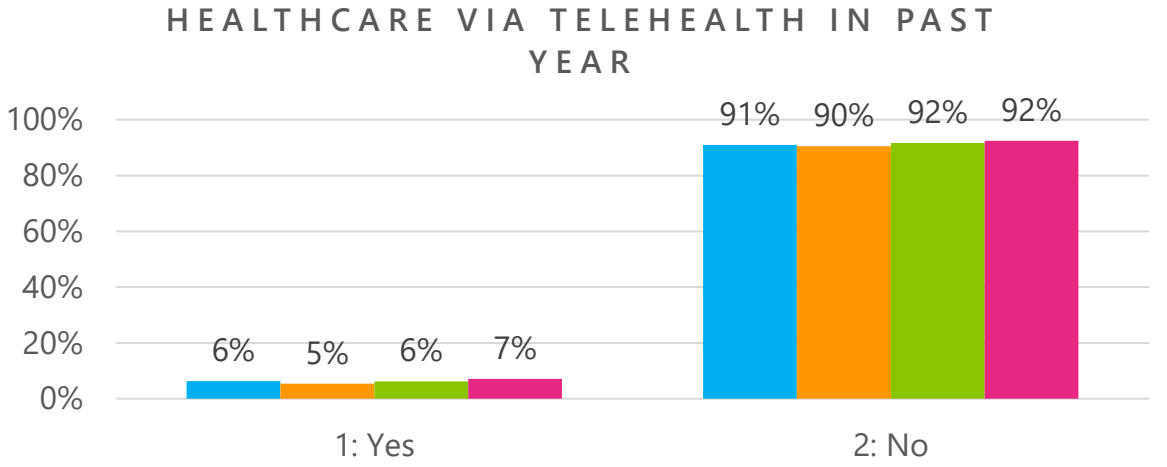




Health Care
Full Sample Estimates

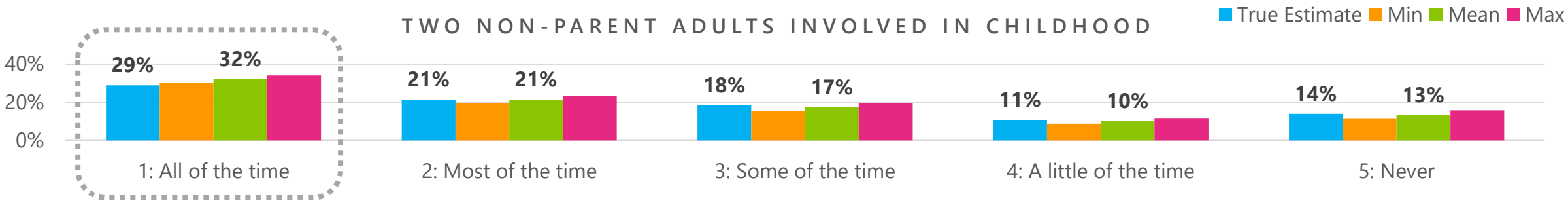


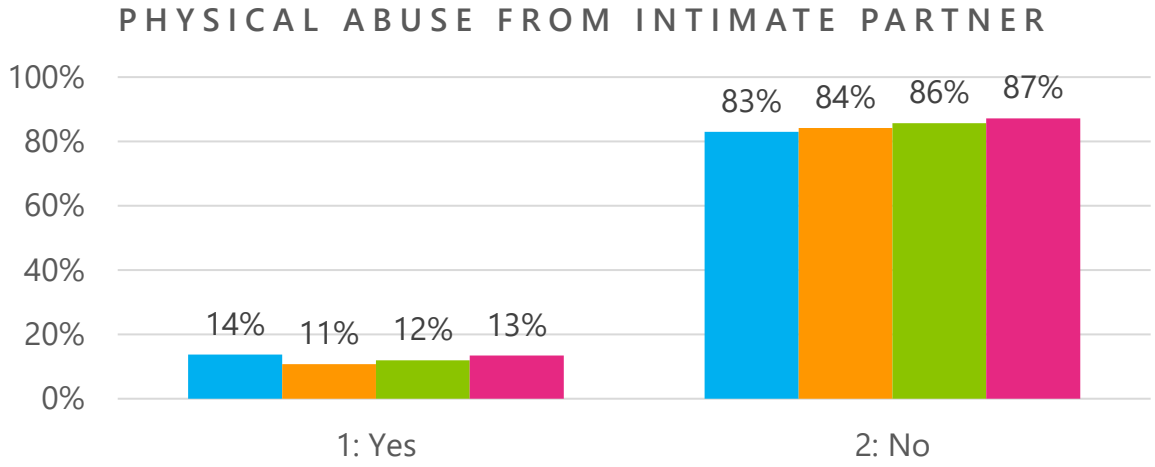


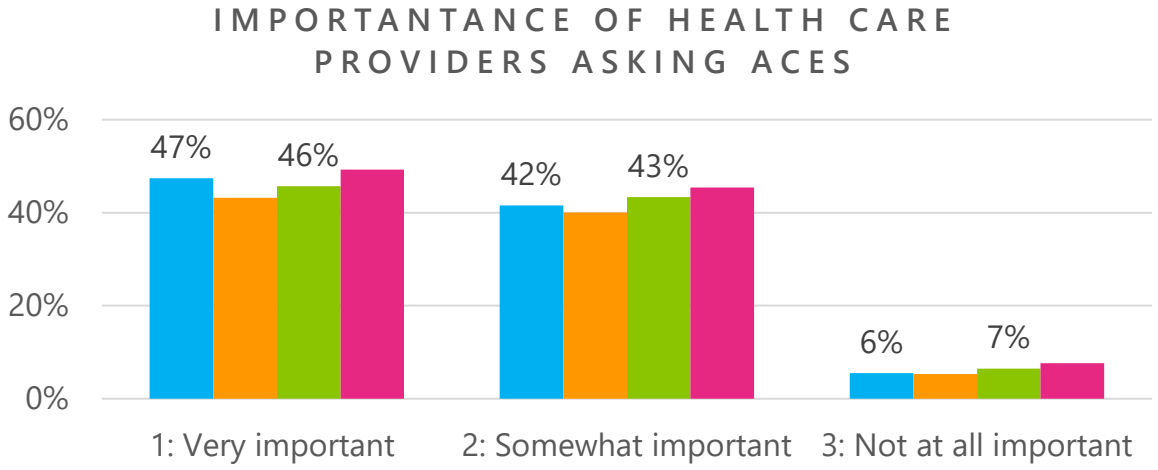


Federal Committee on Statistical Methodology

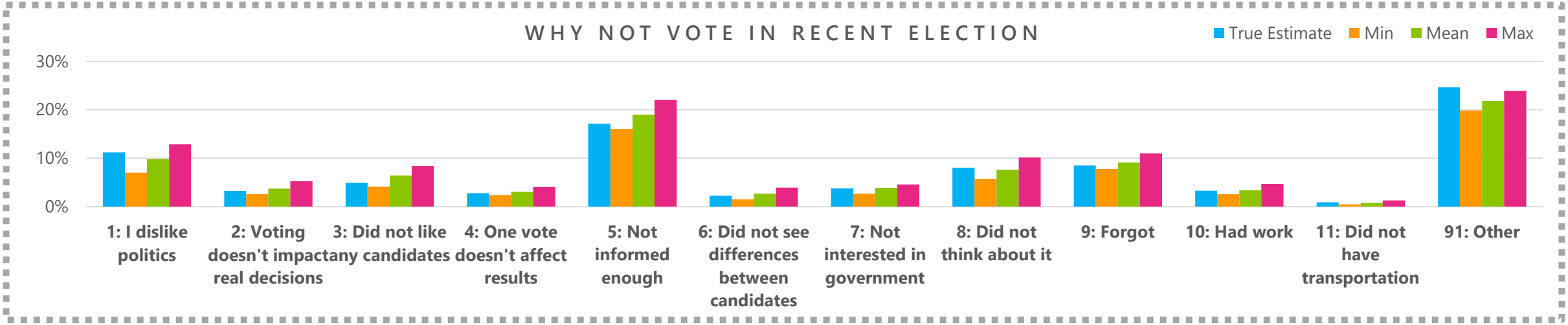
Psychological Distress/Mental Health
Full Sample Estimates

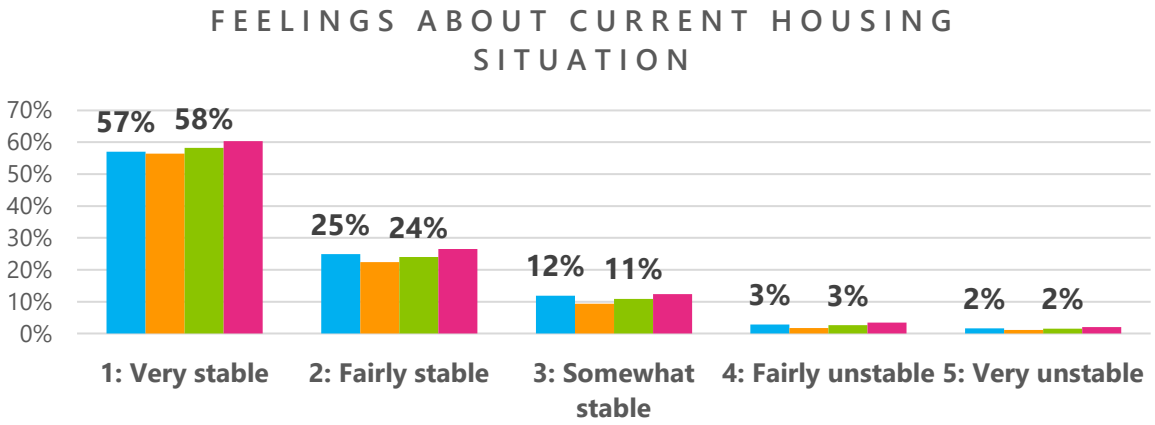


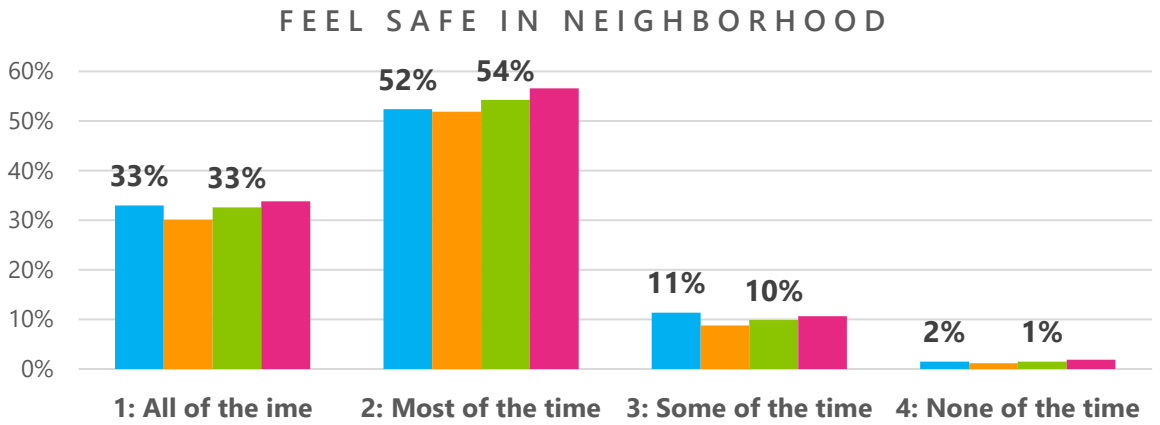




Employment, Housing, & Earnings Full Sample Estimates







Federal Committee on Statistical Methodology

Variances

	PERCENTAGE ESTIMATES	MEAN ESTIMATES
Average True SE/CV	0.304%	2.23%
Average Imputed SE/CV	0.304%	1.94%
Number of SE's smaller than true value	2,226 (53%)	63 (32%)
Number of SE's greater than true value	1,974 (47%)	133 (68%)

Summary

It's HARD!

- Data processing is incredibly time consuming and requires a huge amount of precision coding.

For the outcomes analyzed here, the imputation led to minimal bias on full sample estimates.

- Items with more categories and continuous variables tend to show more likelihood of bias in imputation.

The use of hotdeck imputation did not seem to attenuate variances.

- For continuous variables, the SE on the means seemed to be generally larger for the imputed data (which is not a bad thing).
- This may not hold true for subgroup estimates.

Next Steps

Extend Findings to Additional Outcomes

Overall and to subgroup estimates

Form Reduction

- Are all form pairs necessary?
- Can forms be combined such that the longest and shortest modules are paired to narrow range in timings?

Multivariate Extension

Does the imputation add bias to estimates of relationships?

Add Cyclic Imputation To Increase Precision

Evaluate whether this reduces the differences in min/max and improves the imputation

ssrs
redefining research

THANK YOU

Cameron McPhee

SSRS Chief Methodologist

cmcphee@ssrs.com

Federal Committee on Statistical Methodology