INSTITUTE FOR SOCIAL RESEARCH
POPULATION DYNAMICS AND HEALTH PROGRAM
UNIVERSITY OF MICHIGAN

TOTAL SURVEY ERROR:

A Framework For High-Quality Survey Design

Brady T. West & Paul Schulz Oct. 23, 2018

WORKSHOP OVERVIEW

1. **TSE Introduction & Theory (45 minutes)**
   History, TSE paradigm, & types of error
2. **TSE Measurement, evaluation, and reduction strategies (1 hour)**
   How much error do I have and what can I do about it?
3. **Research Vignette: TSE in practice (45 minutes)**
   The impact of an incentive change on TSE in the NSFG, and future work.
4. **Group Exercise (1 hour)**
   Mock study design challenge!

TOTAL SURVEY ERROR:

Introduction & Theory

Brady T. West & Paul Schulz Oct. 23, 2018

TSE HISTORY

- **Neyman (1934):** Sampling variance properties of descriptive statistics.
- **Deming (1944):** First to outline multiple error sources in surveys ("13 factors").
- **1950s (Hanson, Hurwitz & Madow; Cochran):** Nonsampling error sources acknowledged (Noncoverage, and processing error).
- **Kish (1965):** Focus on bias estimation.
- **Dalenius (1974):** "total survey design"
- **1980s-now (Couper, Biemer, Lyberg, Groves, Heeringa):** Modern concepts of TSE, use of paradata, and total survey quality.

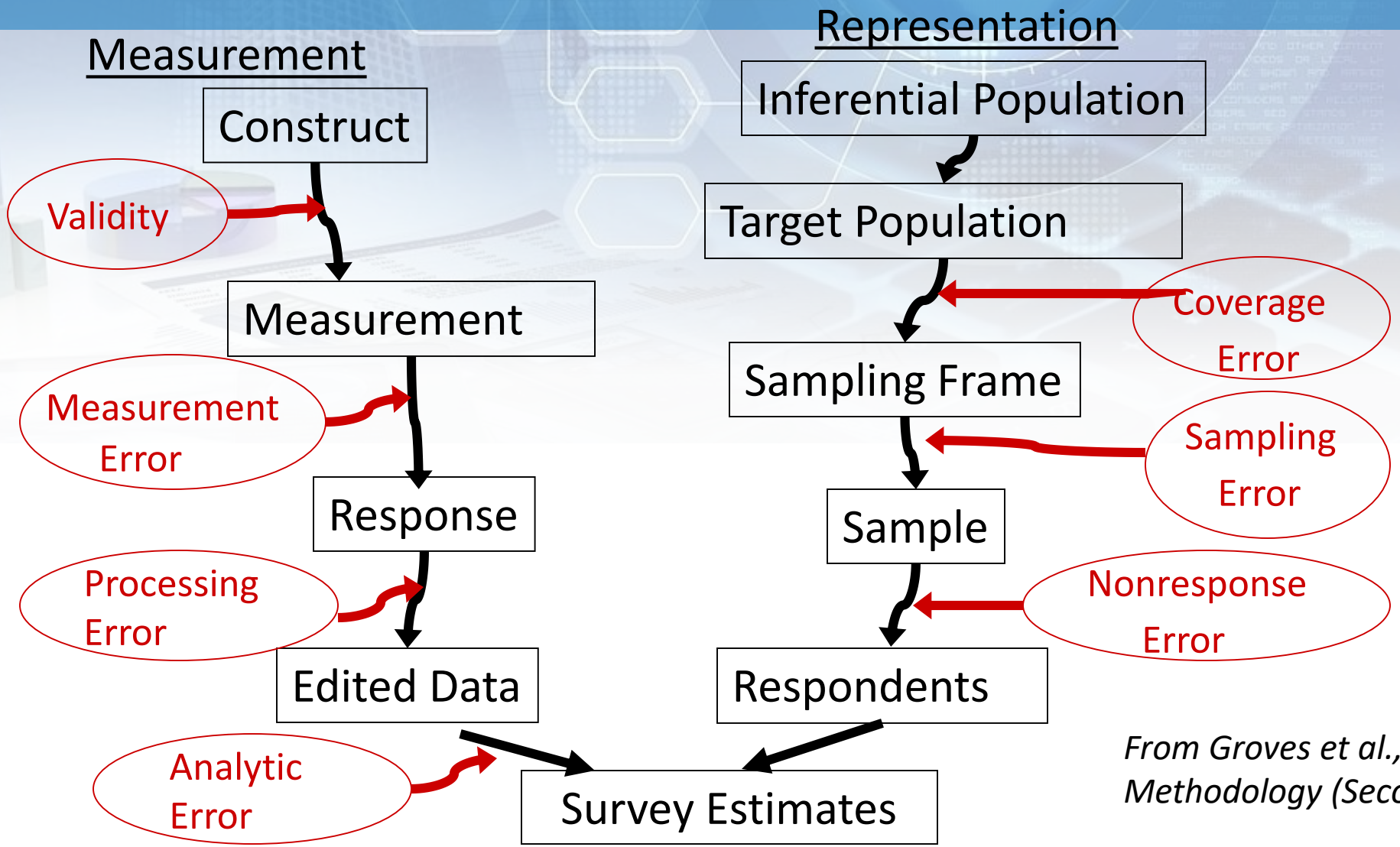
TSE PARADIGM

- Embedded within the broader concept of "total survey quality"
- Total Survey Error paradigm is a guiding framework for optimal allocation of resources to minimize error in key estimates.
- Identifies all major sources of error (which will change over time as technology evolves)
- Seeks to reduce error where possible, while still meeting budget and deadline objectives.

IMPACTS ON STUDY DESIGN

- Study design should take into account all known sources of error, and attempt to reduce/eliminate them where possible.
- Quantification of error can help establish hierarchy of error sources, and set priorities of study team.
- Ultimate goal is not only to meet budget and time constraints, but also produce high-quality data that will produce accurate and robust estimates.
- Keep end data analysts and users in mind.

TSE COMPONENTS



From Groves et al., 2009, Survey Methodology (Second Edition)

TSE COMPONENTS

- Error sources follow a nested structure, with the groups shown above containing smaller and more specific types of error.
- Using Groves et al.'s chart as the canonical example, error sources are split into two parallel branches:

    1. Measurement (errors of construct validity, survey questionnaire design, and processing error), and

    2. Representation (errors of coverage, sample design, and nonresponse)

- Not all texts historically have used this exact same version of the TSE framework, but the basic concepts are always similar.

ERROR SOURCES: VALIDITY

- The gap between a survey measurement and the true value of the underlying construct that is intended to be measured.
- Based on true score theory from the psychometrics field.
- Example: Using the construct of intelligence, the true value could not be captured by a single measure in time or from a single task; it would consist of a more permanent attribute separate from any one single measure.

MEASUREMENT ERROR

- The difference between a specific measured value of a quantity and it's true value.
- Can arise from respondent misreading or misunderstanding a survey question, for example.
- Also known as observational error.
- Measurement error bias: systematic deviation in responses from true values due to poor measurement (shifts estimates!)
- Measurement error variance: variability across measurement occasions / data collection agents in the answers reported

PROCESSING ERROR

- Errors that arise in the transfer of data from the recording of the data to the analytic dataset.
- Examples: misreading of respondent handwriting, data entry errors, errors in scanned questionnaires.
- Common example: open-ended reports of occupations being transferred into occupational codes by human coders
- Are these errors systematic? Or variable across coders?
- Generally receives less focus in our era of CASIC, but can still be an issue depending on the survey!

ANALYTIC ERROR

- Errors that arise in the post-processing steps after data has been collected from the field and stored in an analytic dataset.
- Examples: errors arising from incorrect merging, attribution of response to the wrong individual, incorrect use of survey weights and design features for estimation and inference, etc.
- Unfortunately very common: See West et al. (2016)

COVERAGE ERROR

- Gap between target population and those represented in the sampling frame.
- Example: in phone studies, those without access to a phone number. For address-based sampling, those without a single permanent address.
- Often prisoners (undercoverage) and dorms (undercoverage -- also double-counting with home address).

SAMPLING ERROR

- Error in an estimate that arises due to measuring a sample instead of an entire population.
- This introduces variance into the survey estimates: in theory, the sampling error will vary across repeated samples.
- Generally will be reduced with larger sample sizes.

NONRESPONSE ERROR

- Error that arises from unobserved elements of the selected sample.
- **Unit nonresponse:** When an entire element is not observed (refusal to participate, can't be found, etc.)
- **Item nonresponse:** Single missing measure even when an sampled element is found and participates.
- Introduces bias to survey measures if not missing at random (when unobserved elements differ from observed elements in expectation).

SUMMARY

- The Total Survey Error paradigm is a useful set of principles that guides researchers on sources of error, and the magnitude of those errors.
- Can help researchers make tradeoffs when there are conflicts between cost and time constraints vs. data collection quality.
- Ultimately, paying careful attention to all of the aforementioned error sources will produce better and more useful survey measures and estimates.

TOTAL SURVEY ERROR:

Measurement, Evaluation, & Reduction Strategies

Brady T. West & Paul Schulz Oct. 23, 2018

MEASURING TSE

- Measuring survey error is a key step in the design process, and should factor into design decisions alongside other cost and quality measures.
- Can help minimize error by comparing TSE measures among alternative study designs that are roughly equal on other key dimensions.
- Can also help inform resource allocation within a given design structure by identifying sources of error and allowing cost-benefit analysis of potential solutions.

MEAN SQUARED ERROR

- Most common measure of TSE is Mean Squared Error (MSE):

$$\boldsymbol{MSE}(\hat{\boldsymbol{\theta}}) = \boldsymbol{E}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})^{\boldsymbol{2}}$$

Where $\hat{\theta}$ is a survey estimate, and θ is the true underlying value.

-In words, this is the expected squared difference between an estimate and the underlying parameter being estimated, including *all* sources of error.

-Will always have a positive value; higher values of MSE correspond to more error.

MSE COMPONENTS

- MSE can also be decomposed into a sum of two components:

$$\boldsymbol{MSE}(\hat{\boldsymbol{\theta}}) = \boldsymbol{B}(\hat{\boldsymbol{\theta}})^{\boldsymbol{2}} + \boldsymbol{VAR}(\hat{\boldsymbol{\theta}})$$

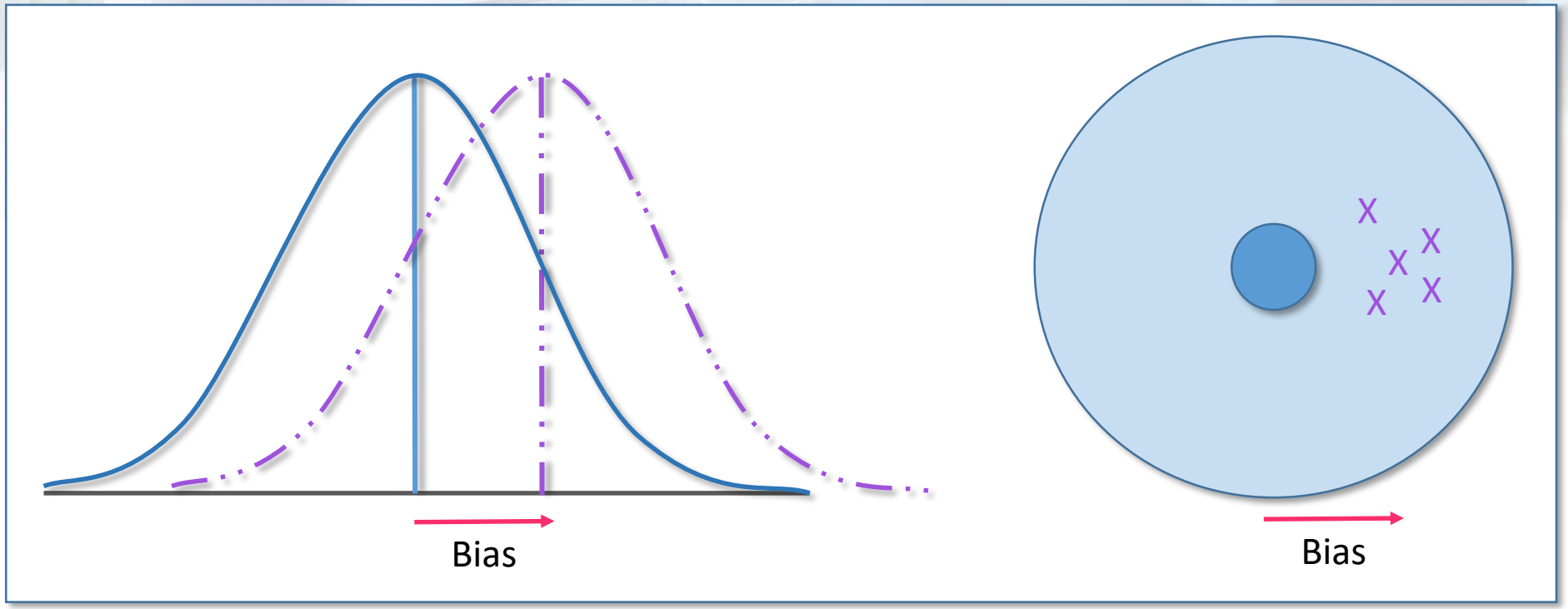
Where $B(\hat{\theta})^2$ is the squared bias of the estimate,
and $VAR(\hat{\theta})$ is the variance of the estimate.

-Each source of error may contribute a systematic error (reflected in the bias component), or a variable error (reflected in the variance), or both.

-Each of these components can in turn be further decomposed into subcomponents.

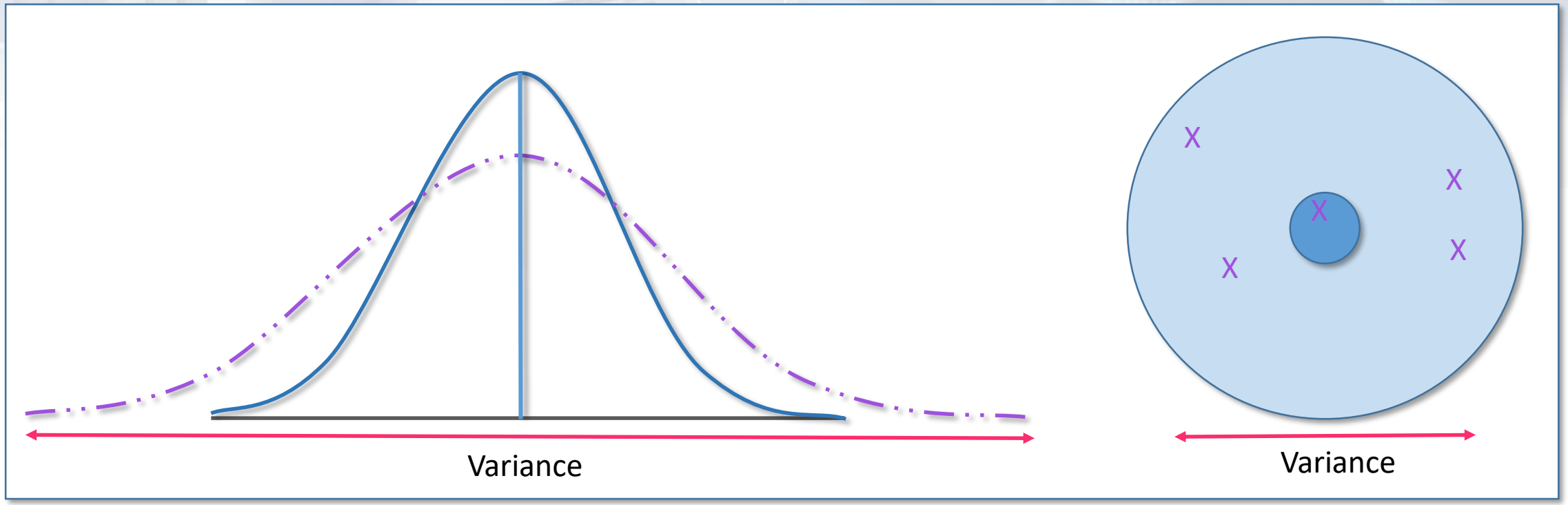
BIAS

Bias is a systematic departure or "shift" of the estimate away from its true value (same precision but decreased accuracy).



VARIANCE

Variance is increased variability of the estimate that is centered around the true value (same accuracy but decreased precision).



TSE EVALUATION

- Just as reporting standard errors is standard practice for post-survey measures, evaluations of non-sampling error sources can also be useful for both study staff and end data users.
- Examples: Non-response bias analyses, measurement reliability studies.
- These evaluations give valuable information to end users about data quality, and help in the understanding and interpretation of results in making inference back to the population of interest.

MSE ESTIMATION

- There is a catch: Calculating MSE directly requires a "gold standard" measurement of underlying parameter(s) of interest θ, which is very rare in practice.
- Gold standard sources: Administrative records (such as census measures, police records, etc.) that are external to the data collection.
- These are often difficult and/or expensive to obtain, and may still be of poor quality.

BIAS ESTIMATION

- If "gold standard" data is available on all respondents and nonrespondents from a given sample, bias is estimated using:

$$\hat{B} = \bar{y} - \mu$$

- Where $\bar{y}$ is the observed sample mean, and μ is the gold standard mean, and assuming a simple random sample.

- Note since $\bar{y}$ and μ are observed from the same sample, any present frame bias will not be reflected in this estimator.

MSE ESTIMATION

- Under the same scenario, an approximate estimator for MSE($\bar{y}$) is given by:

$$\widehat{\text{MSE}}(\bar{y}) = \hat{B}^2 - v(\mu) + 2\sqrt{v(\bar{y})v(\mu)}$$

Where $v(\bar{y})$ and $v(\mu)$ are variance estimators of the sample mean and gold standard mean, respectively.

FRAME BIAS ESTIMATION

- Bias due to frame undercoverage can be estimated if we can obtain estimates of the following two quantities:

  -the non-covered sub-population mean $\bar{y}_{NC}$

  -the proportion of the target population that is uncovered, $\hat{p}_{NC}$

- Frame bias can then be estimated using:

$$\hat{B}_{NC} = \hat{p}_{NC}(\bar{y}_C - \bar{y}_{NC})$$

NONRESPONSE BIAS

- Nonresponse bias can be estimated using a similar formula, again requiring the appropriate estimates of the non-responding elements:

- Nonresponse bias: $\hat{B}_{NR} = \hat{p}_{NR}(\bar{y}_R - \bar{y}_{NR})$

- NOTE: obtaining these estimates can often be costly (both in time and money) and most likely requires a second data collection effort (e.g. the half-open interval method for assessing non-coverage, non-response follow-up study, appending commercial data to a sampling frame, etc.)

BIAS PROPERTIES

- These estimators of bias show that bias is not just a function of the rate of excluded elements, but also the difference between included and excluded elements.
- In other words, if the difference between included and excluded elements is small, then bias may still be small even if the proportion of excluded elements is relatively large.

VARIANCE ESTIMATION

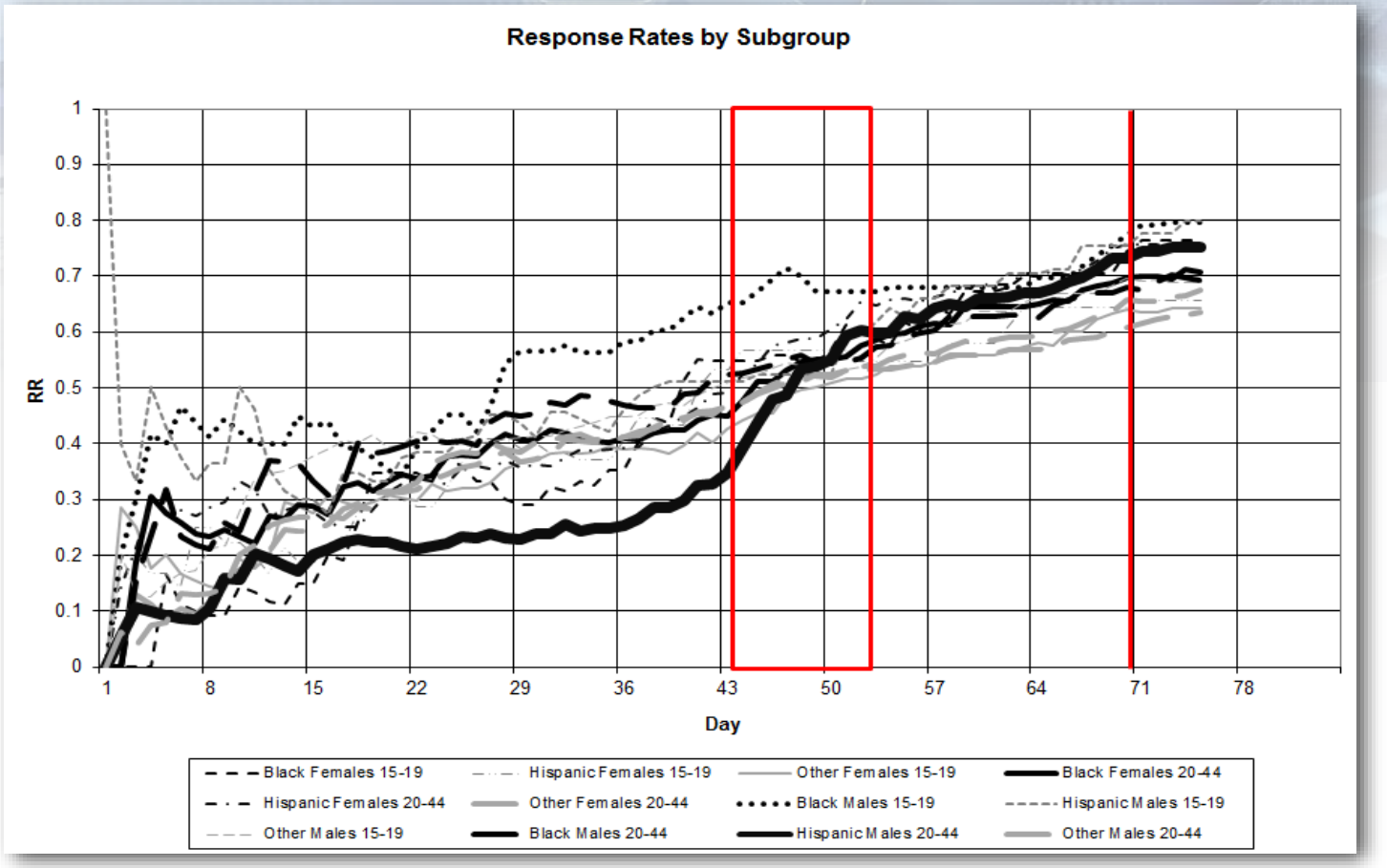
- Sources of variance can also be isolated and measured, again most times requiring a special data collection.
- Example: Interviewer variance estimation (requires randomly assigning cases to interviewers and measuring variance in key measures across interviewers at all stages, including sample assignment, recruitment, and measurement).
- Implications for practice: monitor interviewer variance carefully, and identify unusual interviewers during data collection!

TSE REDUCTION STRATEGIES

- Sources of error that can be identified can be reduced using measurement, monitoring, and allocation of resources.
- Planning involving careful collection of paradata, dashboarding, and/or responsive survey design can greatly aid effectiveness of error reduction techniques. Think ahead!
- Generally speaking, design features of surveys have the property of effect generalizability, so a strategy shown to work in a comparable survey will most likely provide an effective model for your data collection.

TSE REDUCTION EXAMPLES

Example from the NSFG: Reduction of Nonresponse Bias



TSE REDUCTION EXAMPLES

Example from the NSFG: Reduction of Measurement Error

Average of Zscore	Column Labels						
	C8						
	Y2						
	Q01			Q02			
Row Labels	W08	W10	W12	W06	W08	W10	W12
IWERX							
avg_backup_perfield_z	-0.26	-0.35	-0.42	-1.26	-1.41	-1.44	-1.46
avg_DK_perfield_z	1.11	1.37	1.40	0.55	0.77	1.23	1.88
avg_err_esc_perfield_z	-0.38	-0.36	-0.42	-1.12	-1.17	-0.79	-0.74
avg_err_jump_perfield_z	0.93	0.62	0.58	-0.11	0.21	0.70	0.32
avg_err_supp_perfield_z	4.85	4.82	4.86	-0.49	0.18	0.71	0.52
avg_fieldtime_pervisit_z	-0.54	-0.51	-0.49	0.02	-0.06	-0.06	-0.14
avg_qhelp_perfield_z	-0.92	-0.75	-0.77	0.11	0.11	-0.09	-0.19
avg_remclk_perfield_z	2.89	2.91	2.99	2.01	2.58	3.30	3.30
avg_RF_perfield_z	-0.49	-0.42	0.61	-0.30	0.10	0.71	0.50

- Intervene after 2 periods in red
- This interviewer struggled in Q1 in terms of error message suppression
- Intervention appears to work

SUMMARY

-Reducing TSE always requires **advanced planning and early identification of potential error sources**

-Everyone on the survey data collection team needs to be on the same page in terms of a plan to reduce TSE

-Responsive survey design (RSD) can be a helpful tool:

1. Identify potential error sources (e.g., nonresponse bias)
2. Identify indicators of those errors (e.g., subgroup response rates)
3. Develop decision rules for when to intervene / start new phases
4. Monitor the indicators, and intervene / start new phase if decision rule met
5. Combine information across phases / interventions into single estimate
6. Document the results to improve future practice!

RESEARCH VIGNETTE

"The Effects of a Mid-Data Collection Change in Financial Incentives on Total Survey Error in the National Survey of Family Growth: Results from a Randomized Experiment"

-James Wagner, Brady T. West, Heidi Guyer, Paul Burton, Jennifer Kelley, and Mick P. Couper *(University Of Michigan)*

-William D. Mosher *(Johns Hopkins University)*

Originally presented by James Wagner at the 2015 International Total Survey Error conference.

BACKGROUND

Surveys Face Difficulties

- **Declining Response Rates**

  de Leeuw and de Heer, 2002; Brick and Williams, 2013

- **Increasing Effort**

  Curtin, Singer, and Presser, 2005

BACKGROUND

Incentives May Reduce Errors or Costs

- Incentives may increase response rates in face-to-face surveys
  - *Singer and Ye, 2013; Singer et al., 1999; Berlin et al., 1992; Juster and Suzman, 1995; Mosher et al., 1994; Eyerman, et al., 2005*
- Incentives may recruit a "different" set of respondents
  - *Singer and Ye, 2013; Berlin et al., 1992; Juster and Suzman, 1995; Groves, et al., 2005*
- Incentives may reduce costs
  - *Singer and Ye, 2013; Berlin et al., 1992; Kennet and Gfroerer, 2005*
- Incentives may lead to improved reporting
  - *Singer and Ye, 2013; Willimack, et al., 1995; Groves, et al., 2005; Peytchev, Peytcheva, Groves, 2010*

NSFG OVERVIEW

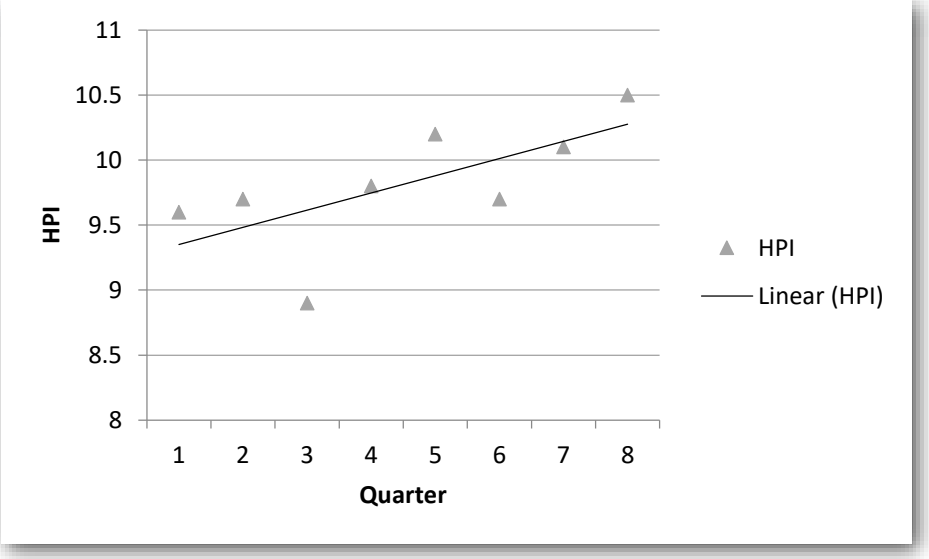
- A study of family life, marriage and divorce, pregnancy, infertility, use of contraception, and men's and women's health
- Multi-stage area sample with face-to-face interviewing
- Continuous design, four 12-week quarters each year

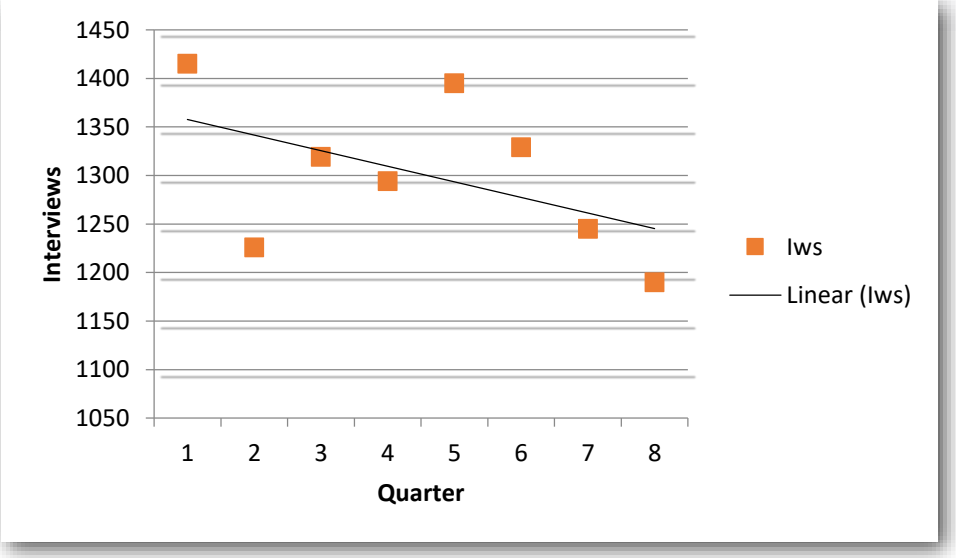
STUDY DESIGN

- **Stages of interviewing:**
  - Screener (identify eligible persons – 15-44 years of age)
  - Main
- **Phases**
  - Phase 1 Weeks 1-10 ($0 for screener, $40 for main)
  - Subsample of nonrespondents
  - Phase 2, Weeks 11-12 ($5 pre-paid for screener, $80 for main)
- **Data**
  - Paradata (call records, timesheets, interviewer observations)
  - Sampling Frame
  - Survey Data

INCENTIVES

- History of experimentation with incentives:
  - 1995: $0 vs. **$20**
  - 2001: $20 vs **$40**
  - 2006: Phase 1: $40. Phase 2 experiment $50 vs **$80**
- 2011-2012: Increasing costs and decreasing yield





EXPERIMENTAL DESIGN

- Phase 1 treatments: $40 (current) versus $60 (experimental)
- Phase 2: $80 for everyone
- Randomization of SSUs to "treatments"
  - Each interviewer has 3
    - Randomized within interviewer
    - Each interviewer has SSUs in each condition
  - 114 SSUs each quarter
- Experiment was run for 5 quarters

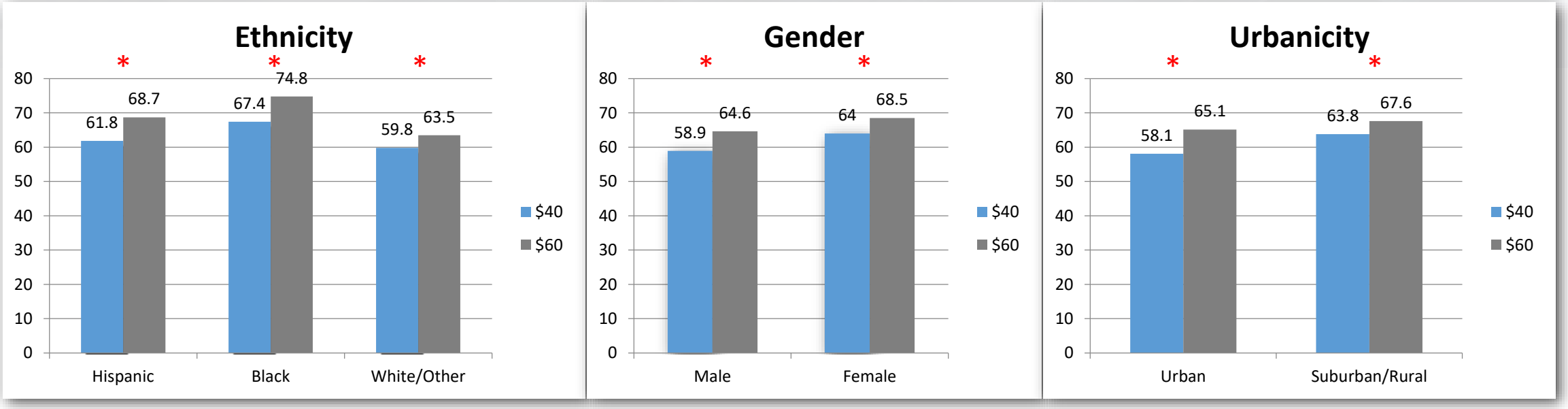
	$40	$60
Sampled Hus	10,966	11,090

RESULTS: Main IW Rate (Phase 1 only)

- Yield Increased:

	$40	$60
Phase 1 Screened Hus	9,205	9,275
Phase 1 Main Iws	2,727	3,139

- Differential impacts on subgroups? (Main Interview Rates)

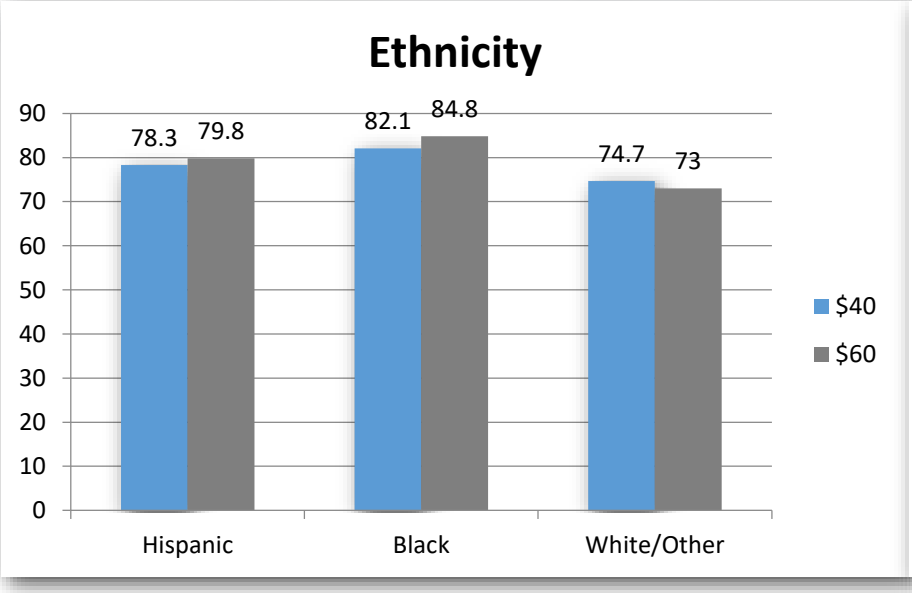


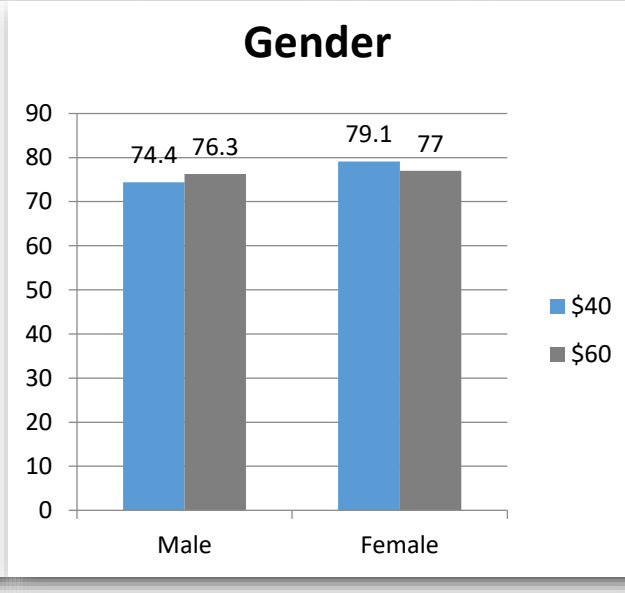
RESULTS: Main IW Rate (Phase 1 & 2)

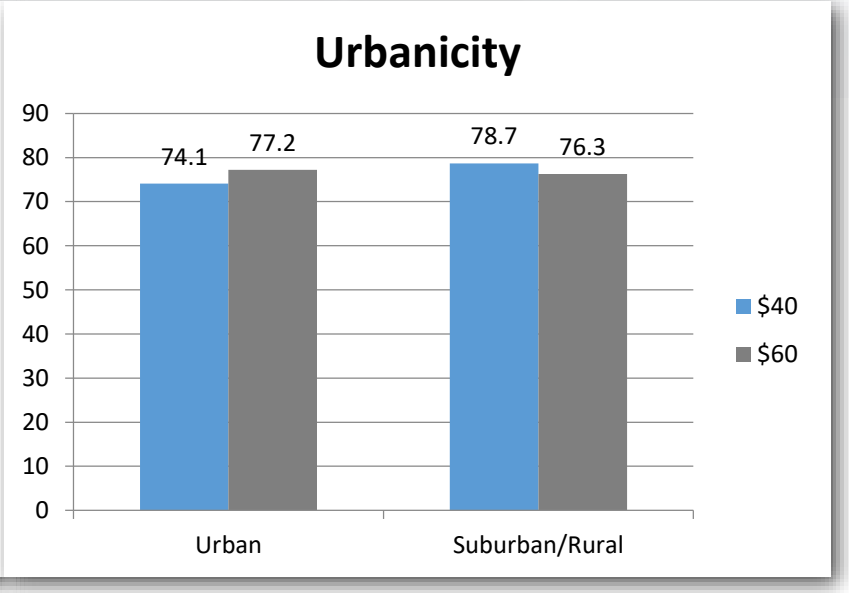
- Yield not Increased:

	$40	$60
Phase 2 Screened Hus	268	294
Phase 2 Main Iws	312	285

- Gains for main interview rate within subgroups are lost:

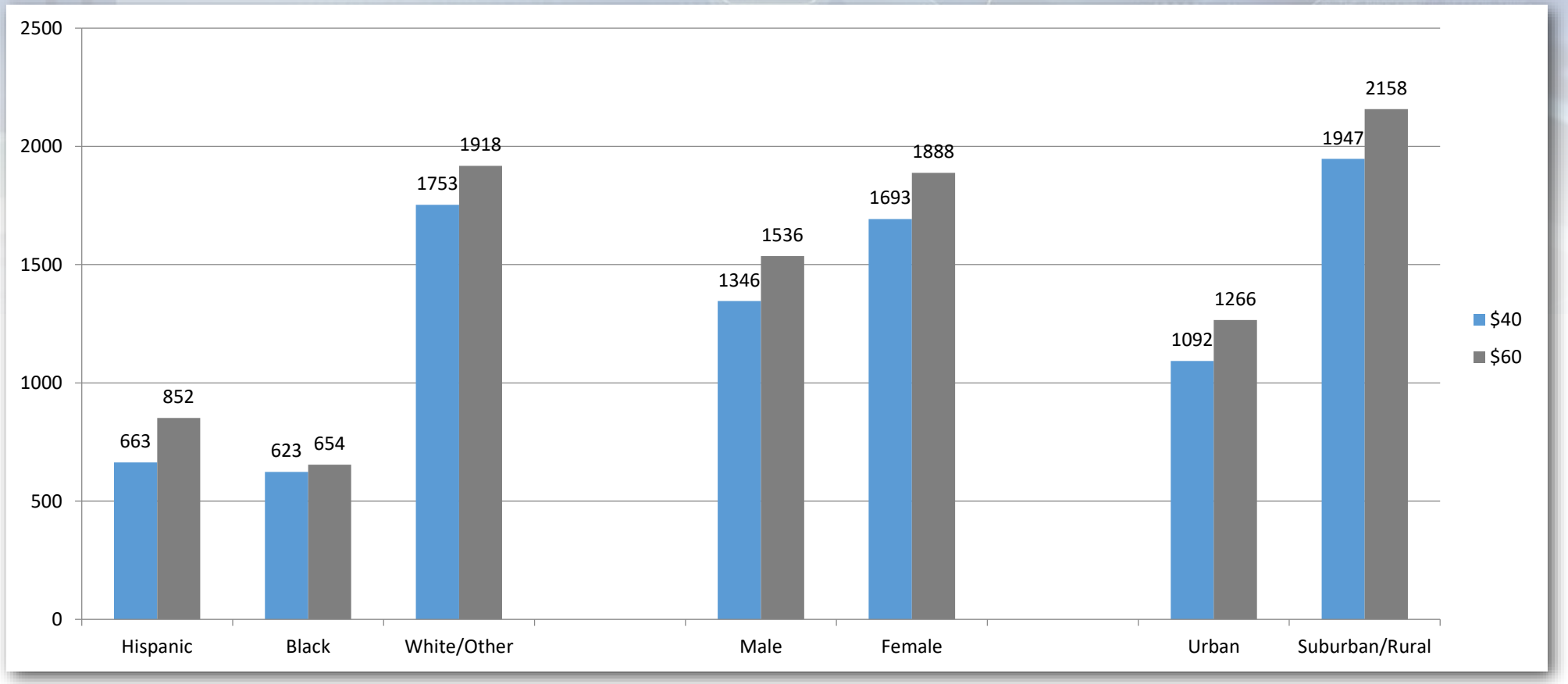






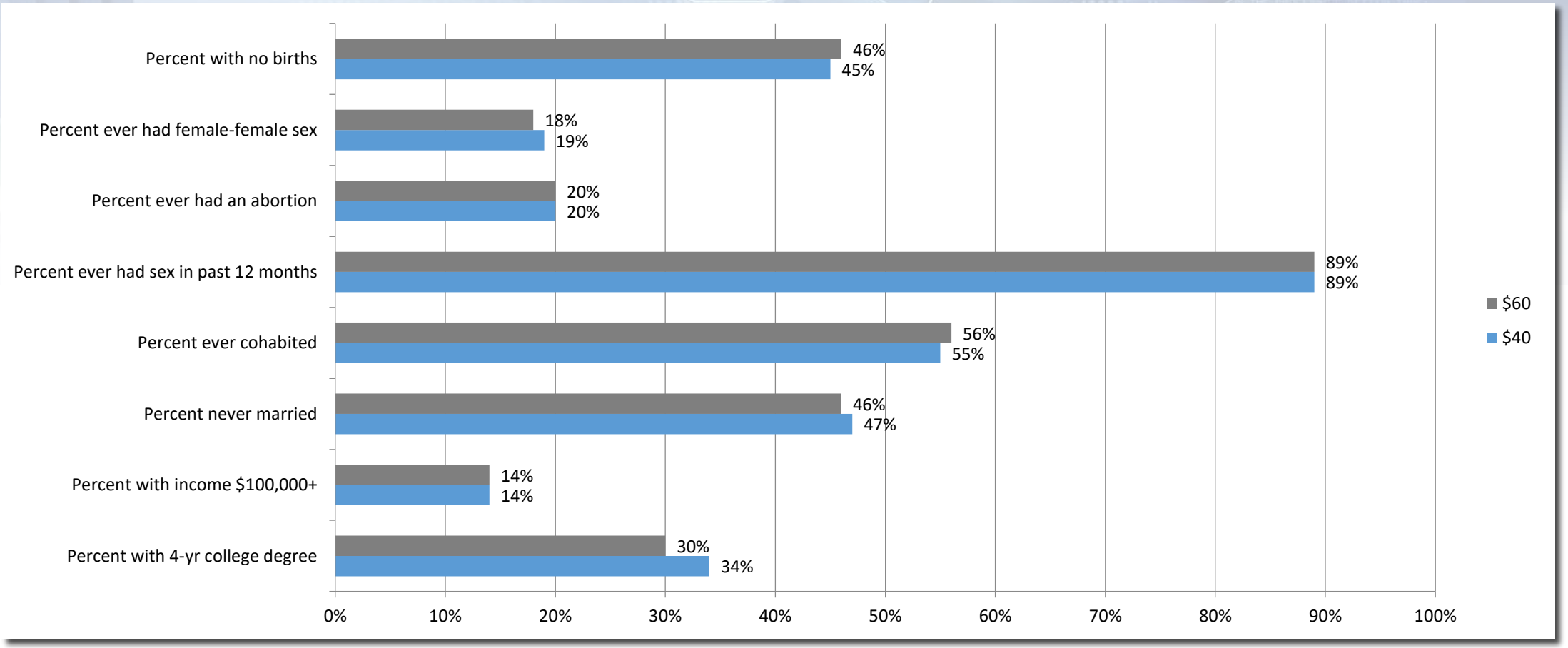
RESULTS: Final Yield (Sampling Error!)

- Yield still higher for subgroups who received $60:



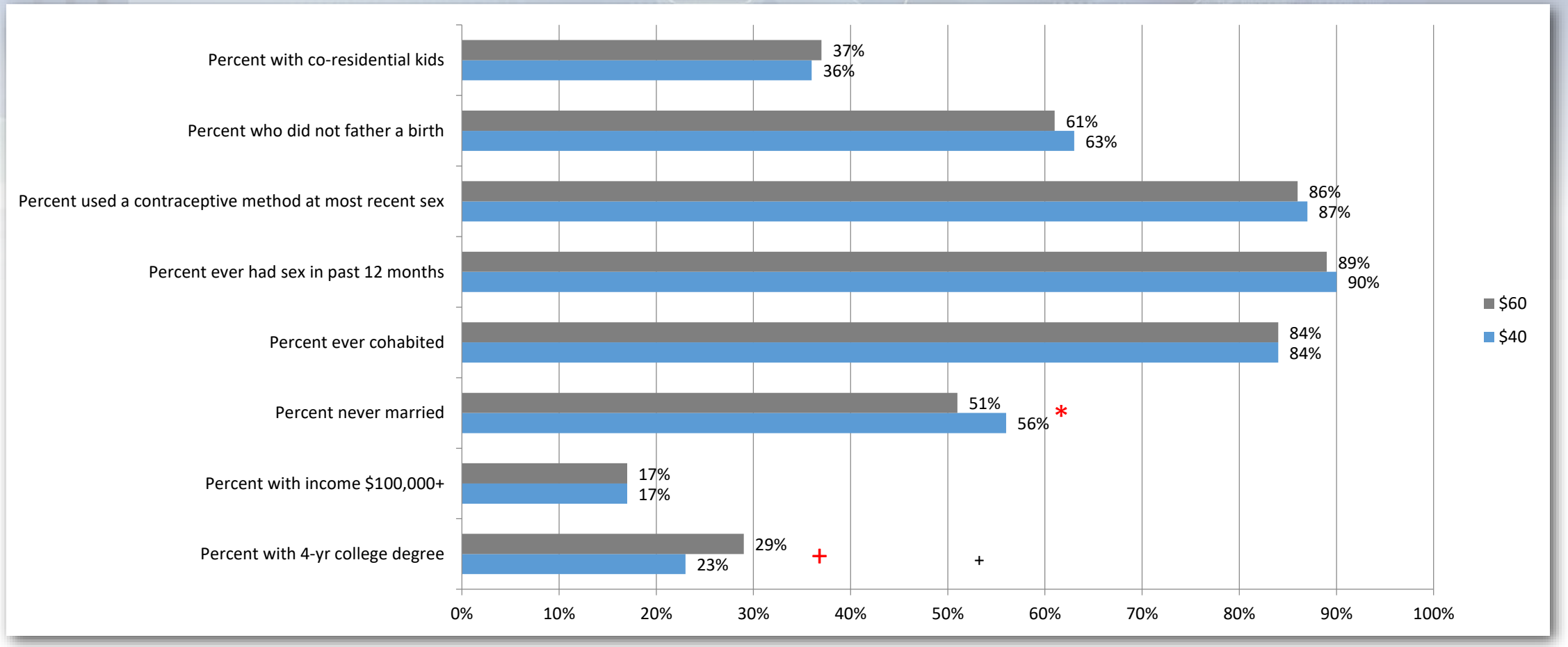
RESULTS: Changes in Estimates

- Females: No changes



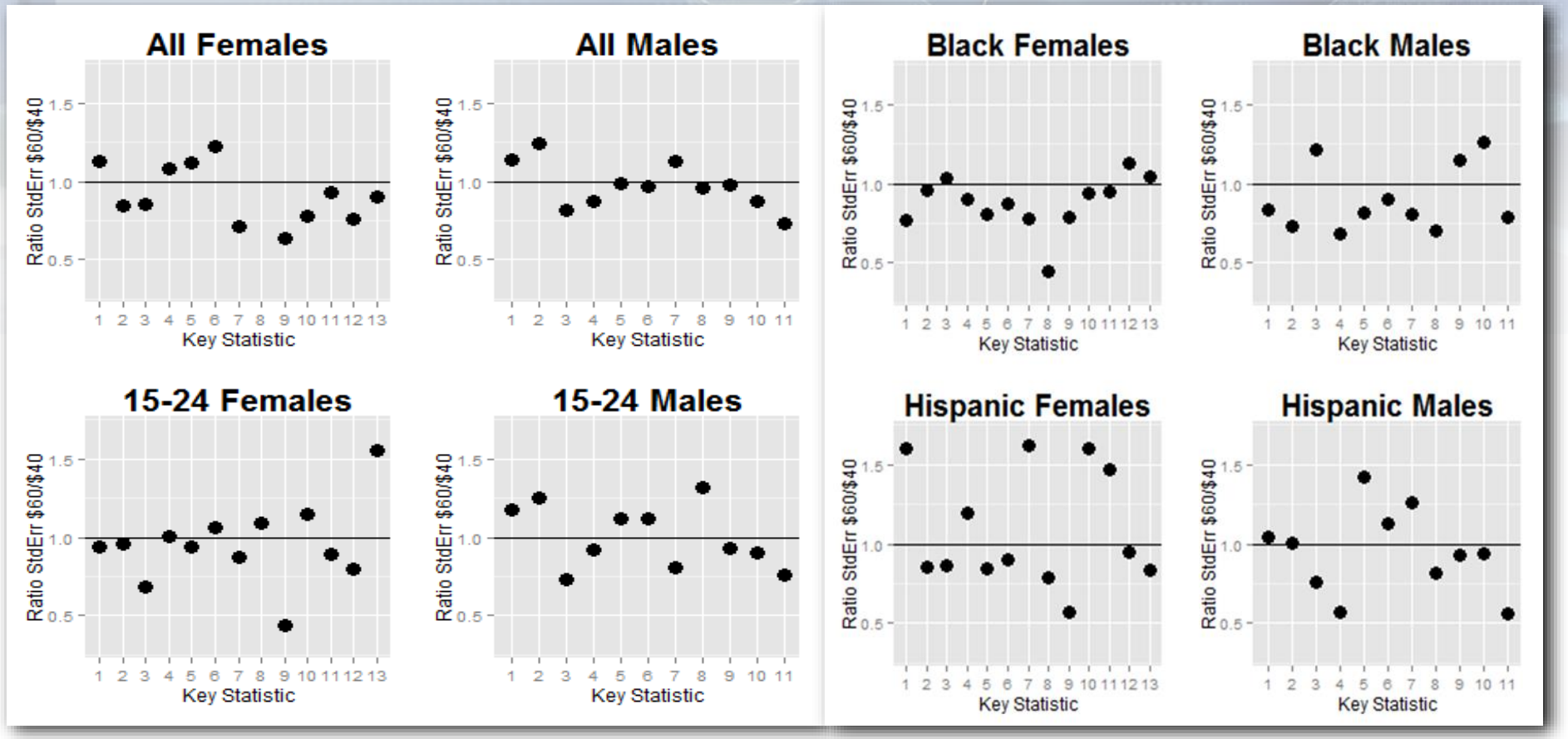
RESULTS: Changes in Estimates

- Males: One difference, one marginal difference



RESULTS: Standard Errors

- For some subgroups, standard errors appear to be smaller



RESULTS: Costs

- **Costs not directly monitored**
  - Interviewers don't report time at the SSU level
  - Need to model costs in some way

- **Indirect measures of effort:**

Calls Per Interview	$40	$60	P-value
Phase 1	26.8	23.4	0.01
Overall	31.9	29.3	0.10

RESULTS: Costs

- **Simple cost model**
  - Average call length=25 minutes
  - Calls saved=2.53
  - Average savings=**63** minutes per interview with $60 (more than $20)
- **More complex cost model**
  - Regression estimates of length of types of calls
  - Estimate: **48** minutes saved per interview with $60
- **Estimated savings: $28 to $37 per interview**

RESULTS: Measurement Error

- Peytchev, Peytcheva, Groves (2010) found that Phase 2 improved reporting
  - Reduced discrepancies between CAPI and ACASI reports of abortion

- **Used ordered logistic regression**
  - -1= ACASI less,
  - 0=equal,
  - 1=ACASI more
- **Covariates:** Age, sex, marital status, phase

No such reduction achieved here

	No Covariates		Covariates	
	Incentive Coefficient	p-value	Incentive Coefficient	p-value
Ever Had Sex (%)	0.0365	0.8871	0.0578	0.8197
Ever Had Sex (Females)	0.2974	0.3940	0.3017	0.3968
Ever Had Sex (Males)	-0.1946	0.5722	-0.1683	0.6148
Mean number of Partners in the Past Year	0.0589	0.5181	0.0715	0.4277
Partners in the Past Year (Females)	0.179	0.0835	0.1803	0.0732
Partners in the Past Year (Males)	-0.0529	0.6946	-0.0261	0.8464
Mean number of Live Births in Past Five Years (Females)	0.00868	0.9071	0.0392	0.6094
Mean number of Pregnancies Fathered in Lifetime (Males)	-0.1059	0.3795	-0.1162	0.3346

SUMMARY

- Increased Phase 1 incentive led to higher yield and higher response rate in Phase 1
  - Especially true for some subgroups: Black, Hispanic, Male, Urban
- Phase 2 results equalized weighted final response rates
- Key estimates not changed by increased incentive
  - Biases not changed
- Sampling error appears to be smaller for increased incentive
  - Especially true for some subgroups
- Measurement error not changed

CONCLUSION

- **The incentive increase accomplished some of its goals**
  - Increased yield (some reduction in sampling error)
  - Decreased cost
- **Other good things did not happen**
  - No evidence of decreased nonresponse bias
  - No evidence of reduced measurement error
- **Incentives may not be useful in reducing all error sources**

CONCLUSION

- **Decided not to switch to $60**
  - Small impact
  - Long-term impact of increasing incentives

- **Why didn't we see the expected effect? *Speculate…***
  - Smaller increase in Phase 2 less effective
    (Would a larger incentive and Phase 1 and 2 be effective?)
  - For some respondents, incentive amount is not the issue

REFERENCES

- Berlin, M., L. Mohadjer, J. Waksberg, A. Kolstad, I. Kirsch, D. Rock and K. Yamamoto (1992). An experiment in monetary incentives. Proceedings of the Survey Research Methods Section, American Statistical Association.
- Brick, J. M. and D. Williams (2013). "Explaining Rising Nonresponse Rates in Cross-Sectional Surveys." The ANNALS of the American Academy of Political and Social Science 645(1): 36-59.
- Curtin, R., S. Presser and E. Singer (2005). "Changes in Telephone Survey Nonresponse over the Past Quarter Century." Public Opinion Quarterly 69(1): 87-98.
- de Leeuw, E. and W. de Heer (2002). Trends in Household Survey Nonresponse: A Longitudinal and International Comparison. Survey Nonresponse. R. M. Groves. New York, John Wiley & Sons: 41-54.
- Eyerman, J., K. Bowman, D. Butler and D. Wright (2005). "The differential impact of incentives on refusals: Results from the 2001 national household survey on drug abuse incentive experiment." Journal of Economic and Social Measurement 30(2): 157-169.
- Groves, R. M., G. Benson, W. D. Mosher, J. Rosenbaum, P. Granda, W. Axinn, J. Lepkowski and A. Chandra (2005). "Plan and operation of Cycle 6 of the National Survey of Family Growth." Vital and health statistics. Ser. 1, Programs and collection procedures(42): 1.
- Juster, F. T. and R. Suzman (1995). "An overview of the Health and Retirement Study." Journal of Human Resources: S7-S56.

REFERENCES

- Kennet, J. and J. C. Gfroerer (2005). Evaluating and improving methods used in the National Survey on Drug Use and Health. Rockville, MD:, Department of Health and Human Services, Substance Abuse and Mental Health Services Administration, Office of Applied Studies.
- Mosher, W. D., W. F. Pratt and A. P. Duffer (1994). "CAPI, Event Histories, and Incentives in the NSFG Cycle 5 Pretest." Proceedings of the Section on Survey Research Methods, American Statistical Association, Toronto.
- Peytchev, A., E. Peytcheva and R. M. Groves (2010). "Measurement Error, Unit Nonresponse, and Self-Reports of Abortion Experiences." Public Opinion Quarterly 74(2): 319-327.
- Singer, E., J. V. Hoewyk, N. Gebler, T. Raghunathan and K. McGonagle (1999). "The Effect of Incentives on Response Rates in Interviewer-Mediated Surveys." Journal of Official Statistics 15(2): 217-230.
- Singer, E. and C. Ye (2013). "The Use and Effects of Incentives in Surveys." The ANNALS of the American Academy of Political and Social Science 645(1): 112-141.
- Willimack, D. K., H. Schuman, B.-E. Pennell and J. M. Lepkowski (1995). "Effects of a Prepaid Nonmonetary Incentive on Response Rates and Response Quality in a Face-to-Face Survey." The Public Opinion Quarterly 59(1): 78-92.

TOTAL SURVEY ERROR:

Mock Study Design Challenge

Brady T. West & Paul Schulz Oct. 23, 2018

STUDY DESIGN CHALLENGE

- **Mock Study Objectives:**

  Nationally representative study of young adults (age 18-35) that will measure their outlook of the future, including estimates of differences in optimism based on current status (socioeconomic, health, etc.), including some demographic subgroup analysis.

3 key survey constructs to measure:

1. Demographics including your subgroups of interest
2. Current status of the respondent (again based on your key status subgroups of interest).
3. Respondent opinions of their future prospects.

STUDY DESIGN CHALLENGE

Key Decisions for your group (be ready to report back!):

- **Mode:** Face to face, phone, web, other?
- **Sample:** Potential frames, sampling strategies (oversample of your key demographic subgroup?).
- **Error Profile** (refer to your handout): Noncoverage error, potential nonresponse bias, validity, measurement error, etc.
- **TSE reduction strategies:** What are key paradata variables, how will they be measured, and how will they be monitored? Is an experiment or meta-study necessary?
- **Resources for the end data user** about sources and magnitude of error in final survey measures.

Good luck! May the best team win!

CLOSEOUT

Thank you for your attendance!

We would love your feedback. Please fill out our "exit survey" or reach out to us with your comments.

- Contact:

Paul Schulz: pschulz@umich.edu, 2012 ISR, 763-2324

Brady T. West: bwest@umich.edu, www.umich.edu/~bwest