Paradata and Visualization

Andrew Mercer
JPSM, Westat

Frauke Kreuter
JPSM, IAB

February 10, 2014

Features of Paradata

- Data that is a byproduct of the data collection process (Couper, 1998)
  - Records of call
  - Response times
  - Keystrokes
- More recently has come to include:
  - GIS data about interviewer location
  - CARI recordings of survey interactions
  - Interviewer observations
- Becoming less incidental and more intentional

Things We Need To Keep In Mind

- Extremely variable in quality and messy
  - What counts as a contact attempt?
- Often an inadequate proxy for what we really want to measure
  - Response times and mouse movements in web surveys as a proxy for difficulty
- Hard to analyze statistically
  - Often requires extensive data processing to be usable
  - Large number of observations challenge decisions based on statistical significance

Paradata and Adaptive Design

- Using data gathered prior to or during data collection to tailor data collection procedures to individual respondents
- Usually optimizing on cost or quality
- Two main uses for paradata:
  - Historical data can be used for analysis and planning of future data collection
  - Current data can be used for monitoring and intervening

Why Visualization?

- Humans are very good at finding patterns in visually presented data
  - (Although sometimes we're too good)
- Graphics are excellent for communicating quantitative information to non-statisticians
  - (e.g. survey managers, interviewers, clients, funders)
- Much of the time you don't actually need fancy models (although those can be fun too)

We See A Lot of Survey Reports that Look Like This:

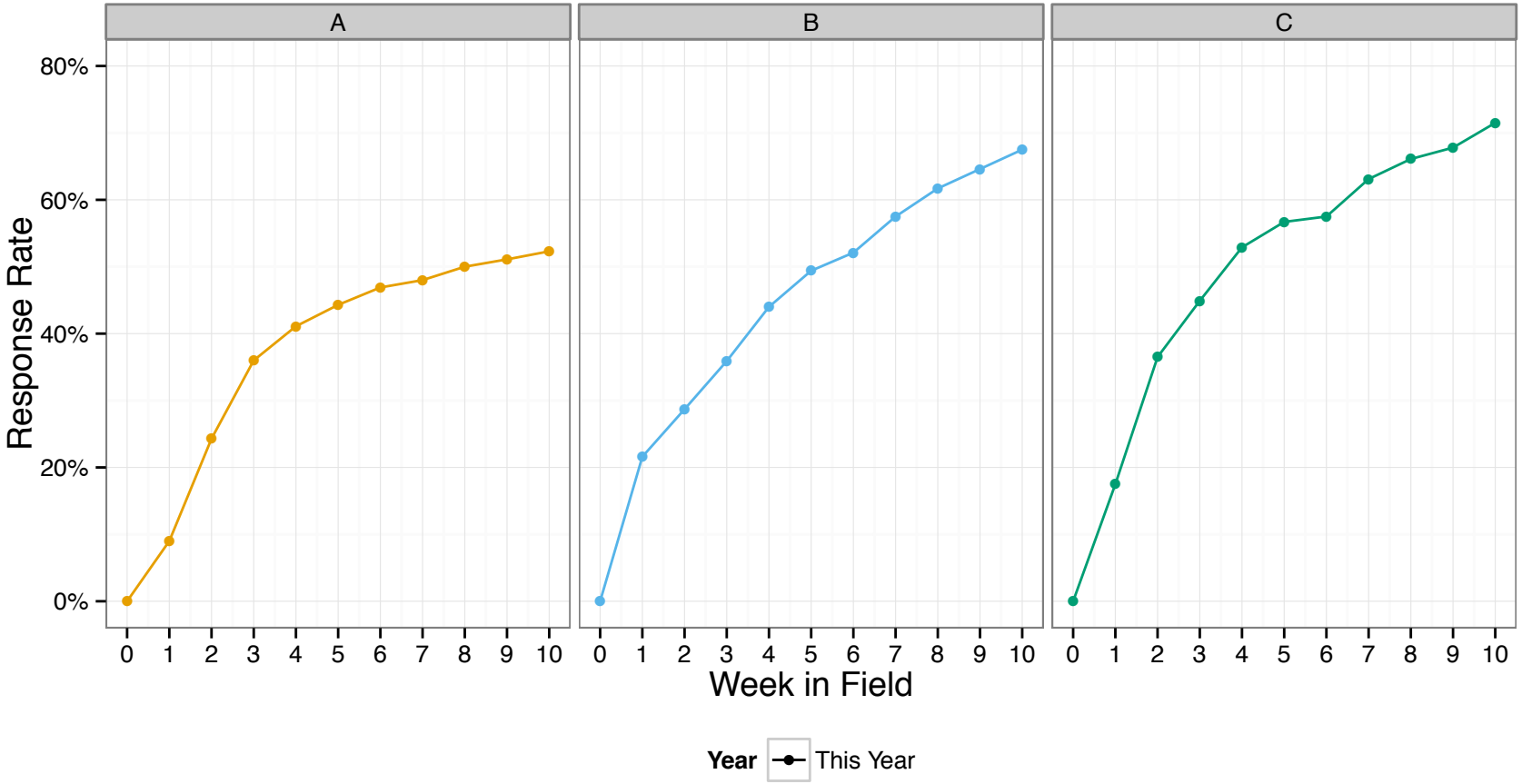
Report # 375 – Cumulative Response Rate This Year vs. Last Year By Region and Field Week

Region A	Week	1	2	3	4	5	6	7	8	9	10
	This Year	9.0%	24.3%	36.0%	41.1%	44.3%	46.9%	48.0%	50.0%	51.1%	52.3%
	Last Year	20.2%	35.5%	47.2%	52.3%	55.5%	58.1%	59.2%	61.2%	62.3%	63.5%

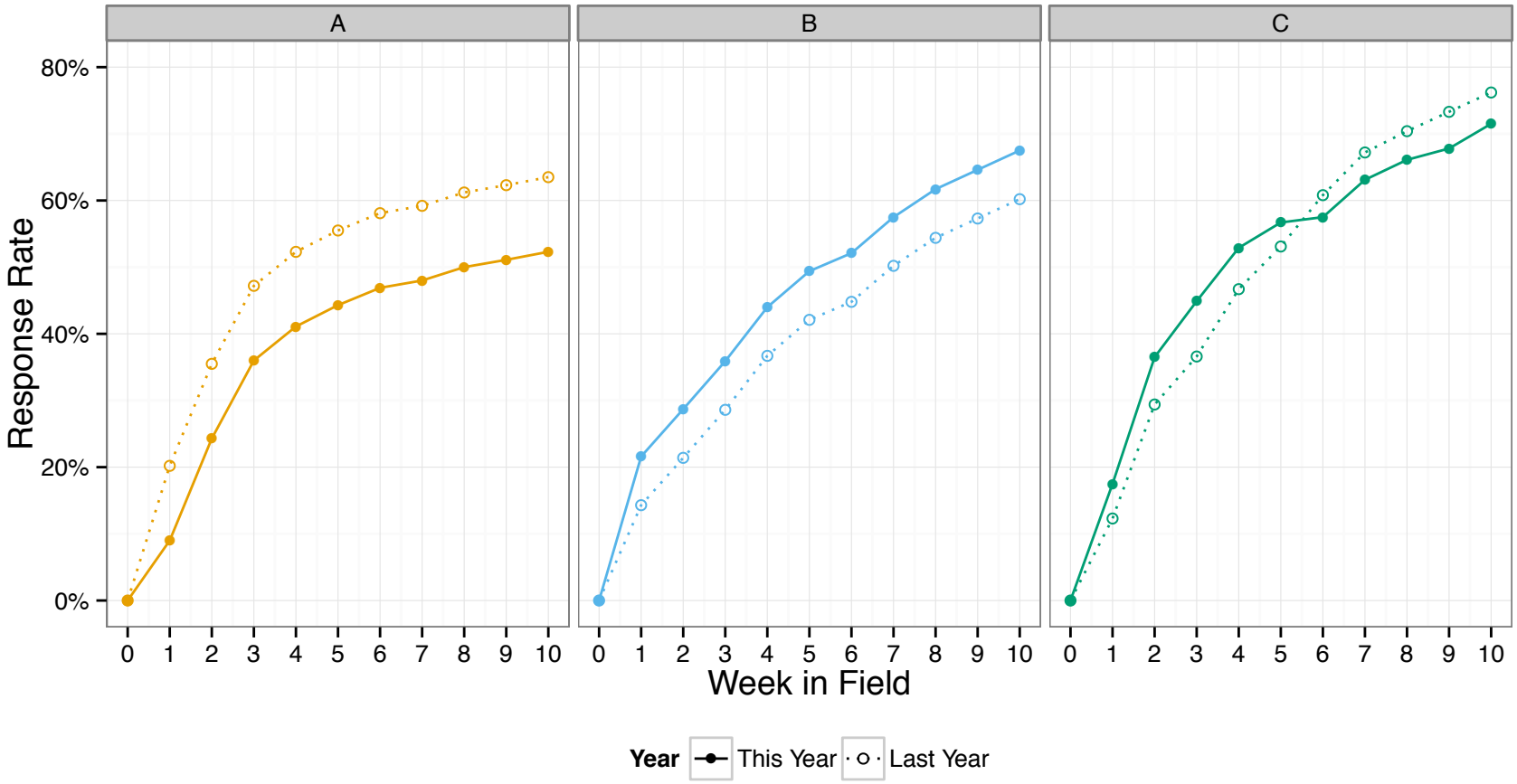
Region B	Week	1	2	3	4	5	6	7	8	9	10
	This Year	21.6%	28.7%	35.9%	44.0%	49.4%	52.1%	57.5%	61.7%	64.6%	67.5%
	Last Year	14.3%	21.4%	28.6%	36.7%	42.1%	44.8%	50.2%	54.4%	57.3%	60.2%

Region C	Week	1	2	3	4	5	6	7	8	9	10
	This Year	17.5%	36.5%	44.9%	52.9%	56.7%	57.5%	63.1%	66.1%	67.8%	71.5%
	Last Year	12.3%	29.4%	36.6%	46.7%	53.1%	60.8%	67.2%	70.4%	73.3%	76.2%

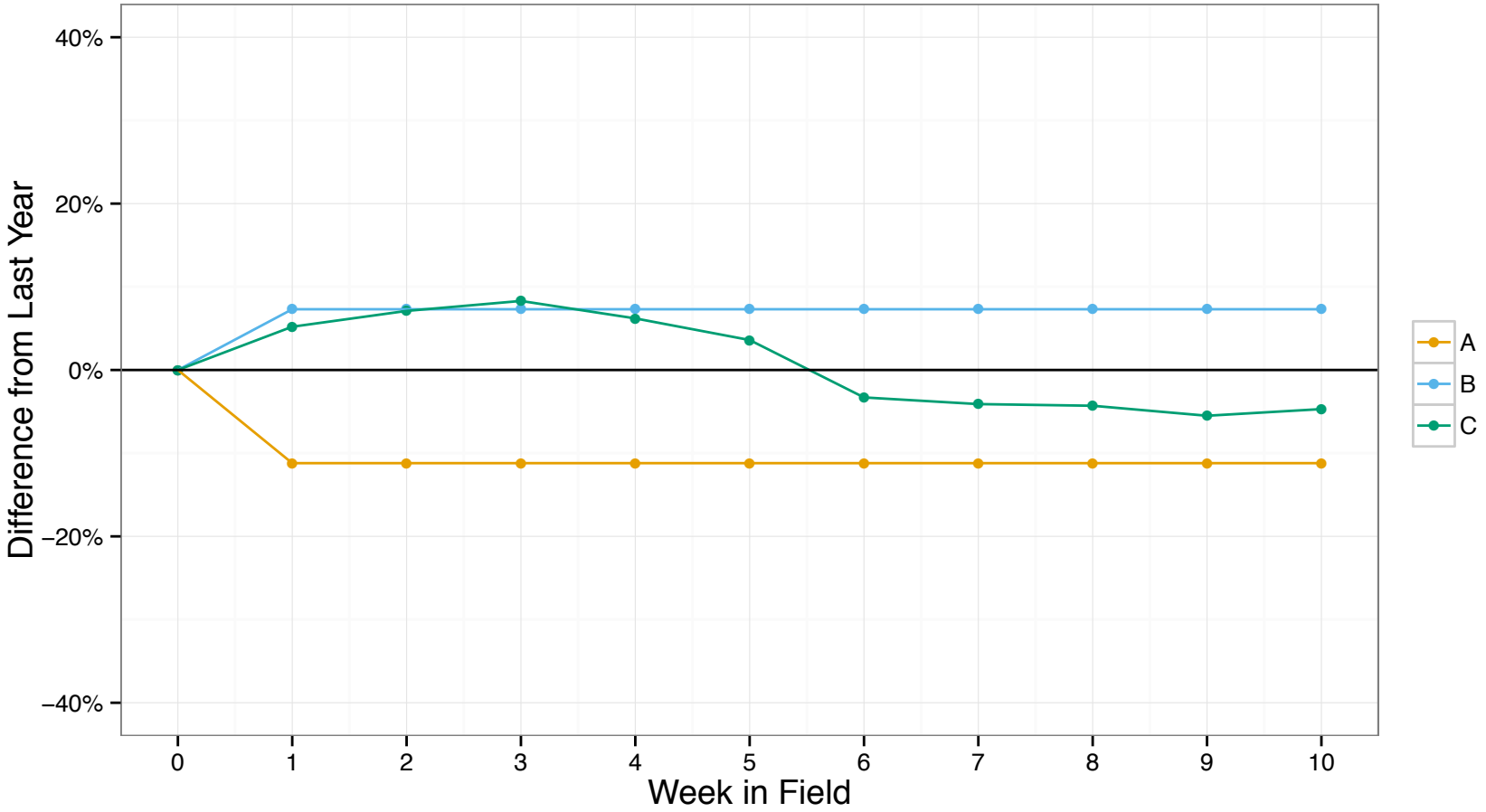
Quickly Spot Patterns and Trends



Draw Comparisons to Last Year



Hone In On the Key Comparisons



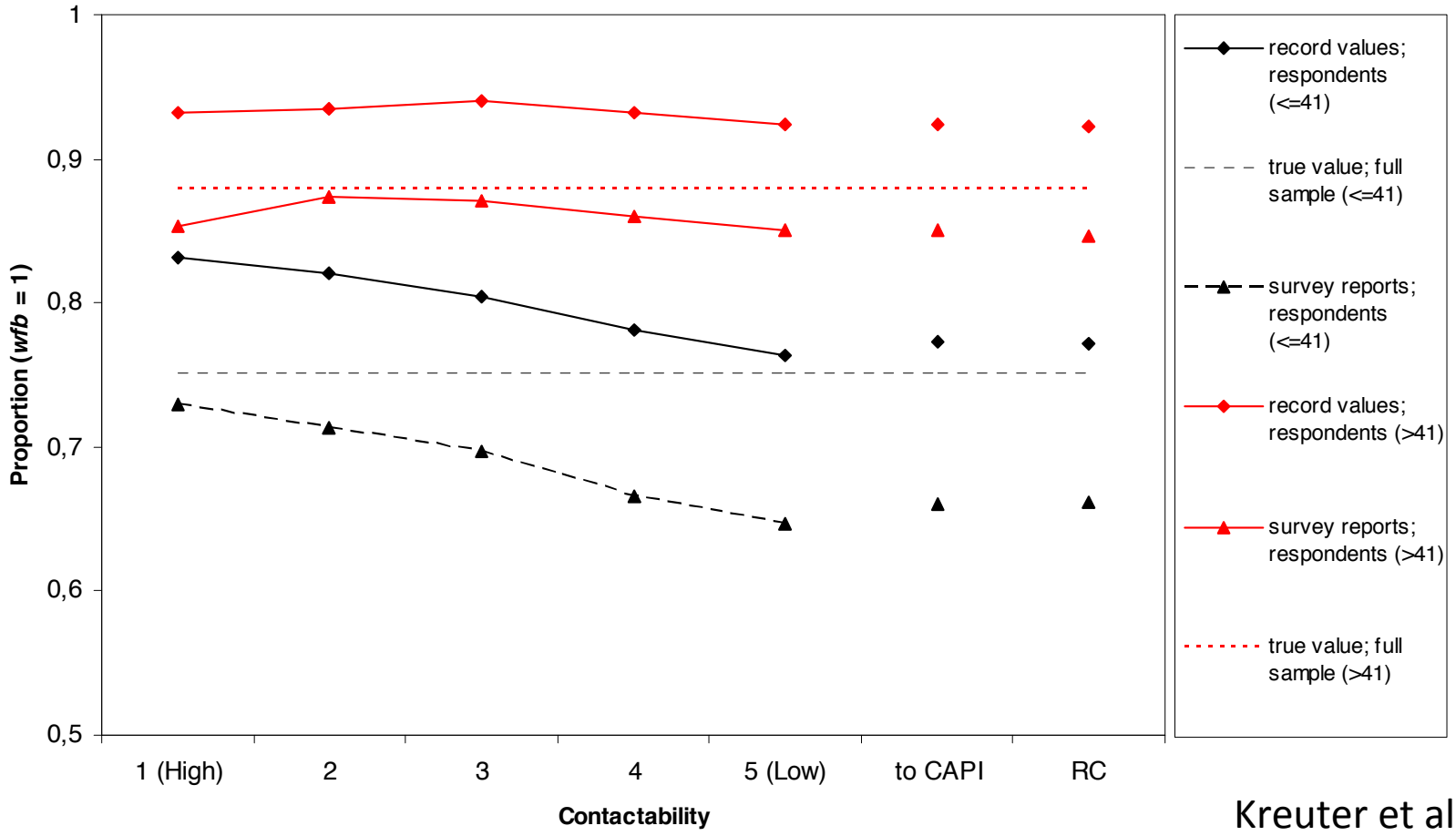
ANALYSIS AND PLANNING

Paradata for Analysis and Planning

- "What-if" Scenarios and Simulation
  - What if we: shortened the field period, set limits on contact attempts, etc…
- Pareto Analysis
  - Identify major drivers of cost, error, etc…
  - Help set priorities for intervention

What-If Scenarios
NR and Measurement Error in PASS

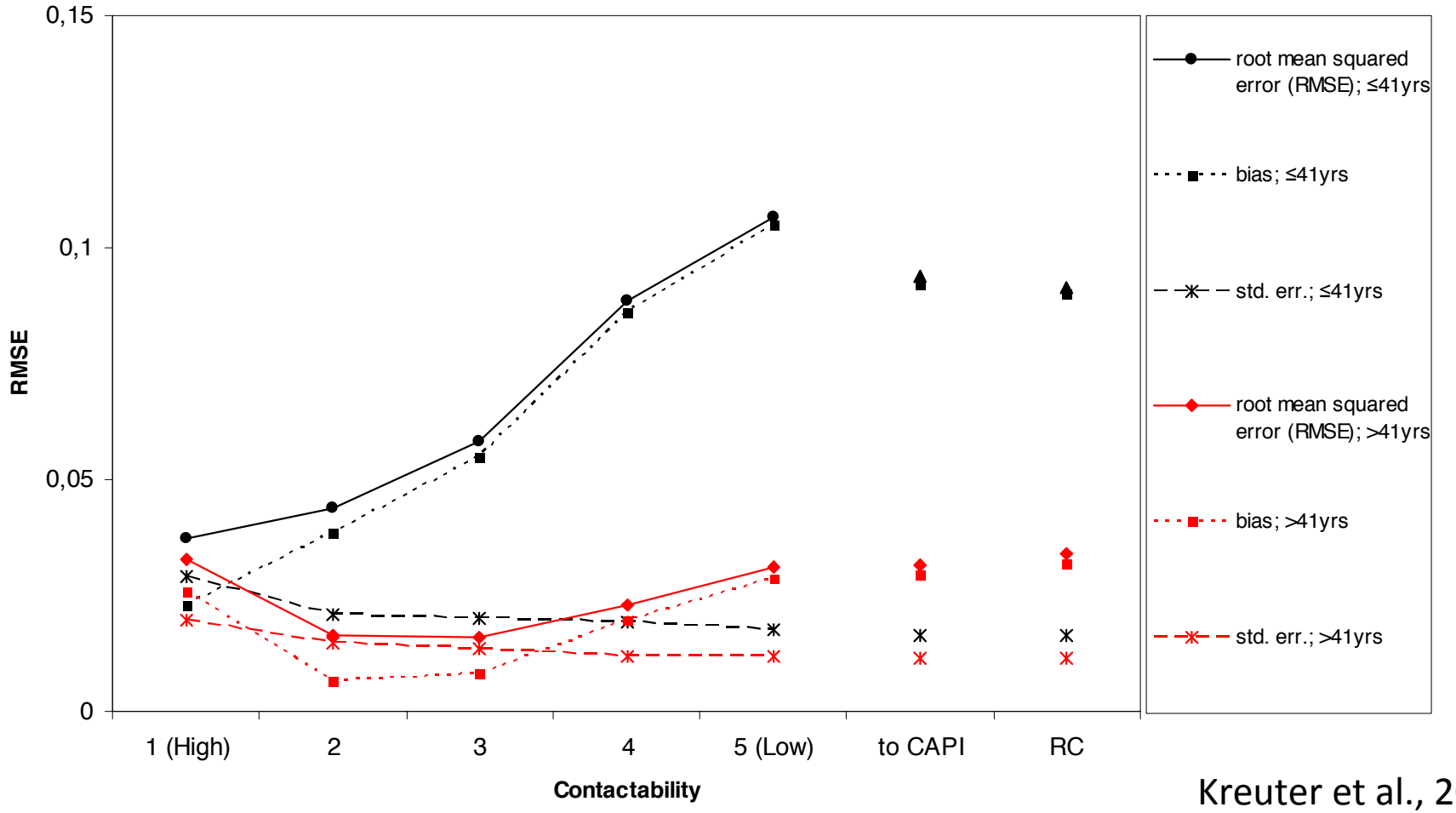
Figure 1a* Cumulative mean over quintiles of no. of contact attempts; Current Welfare Status, by age group



Kreuter et al., 2010

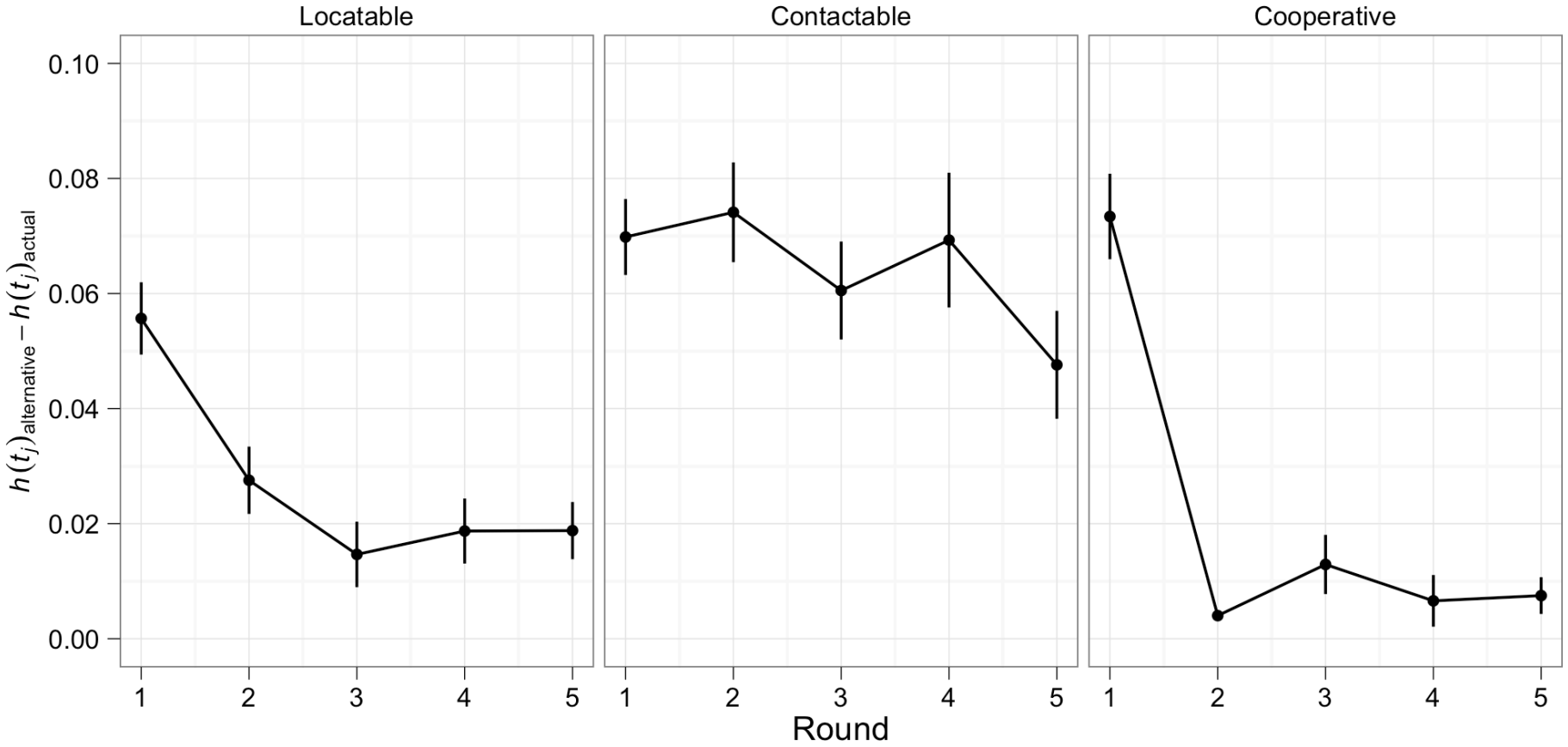
What-If Scenarios MSE

Figure 1b* RMSE over quintiles of no. of contact attempts; Current Welfare Status (*wfb*); by age



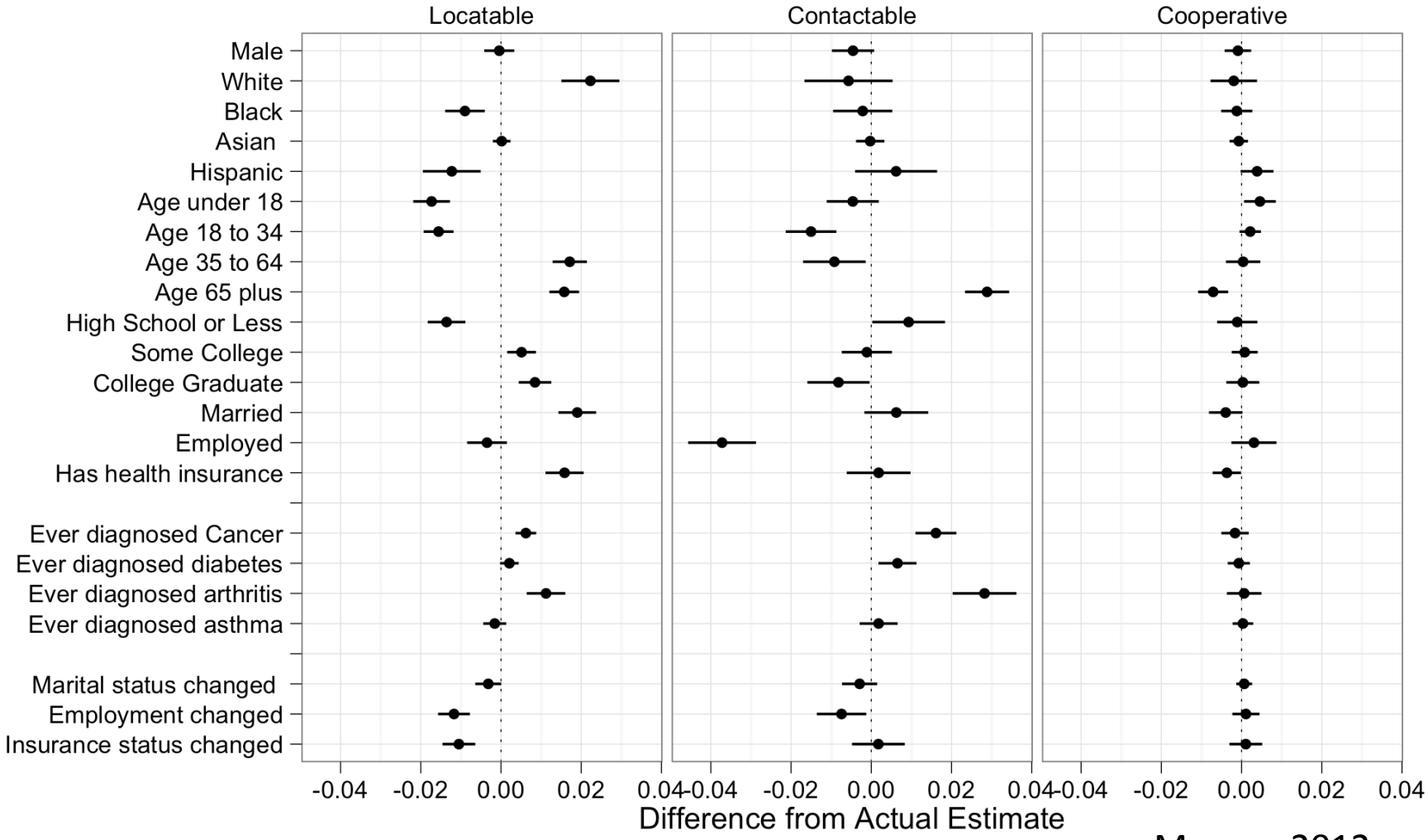
Kreuter et al., 2010

What-If Scenarios
Survey Effort and Attrition in MEPS



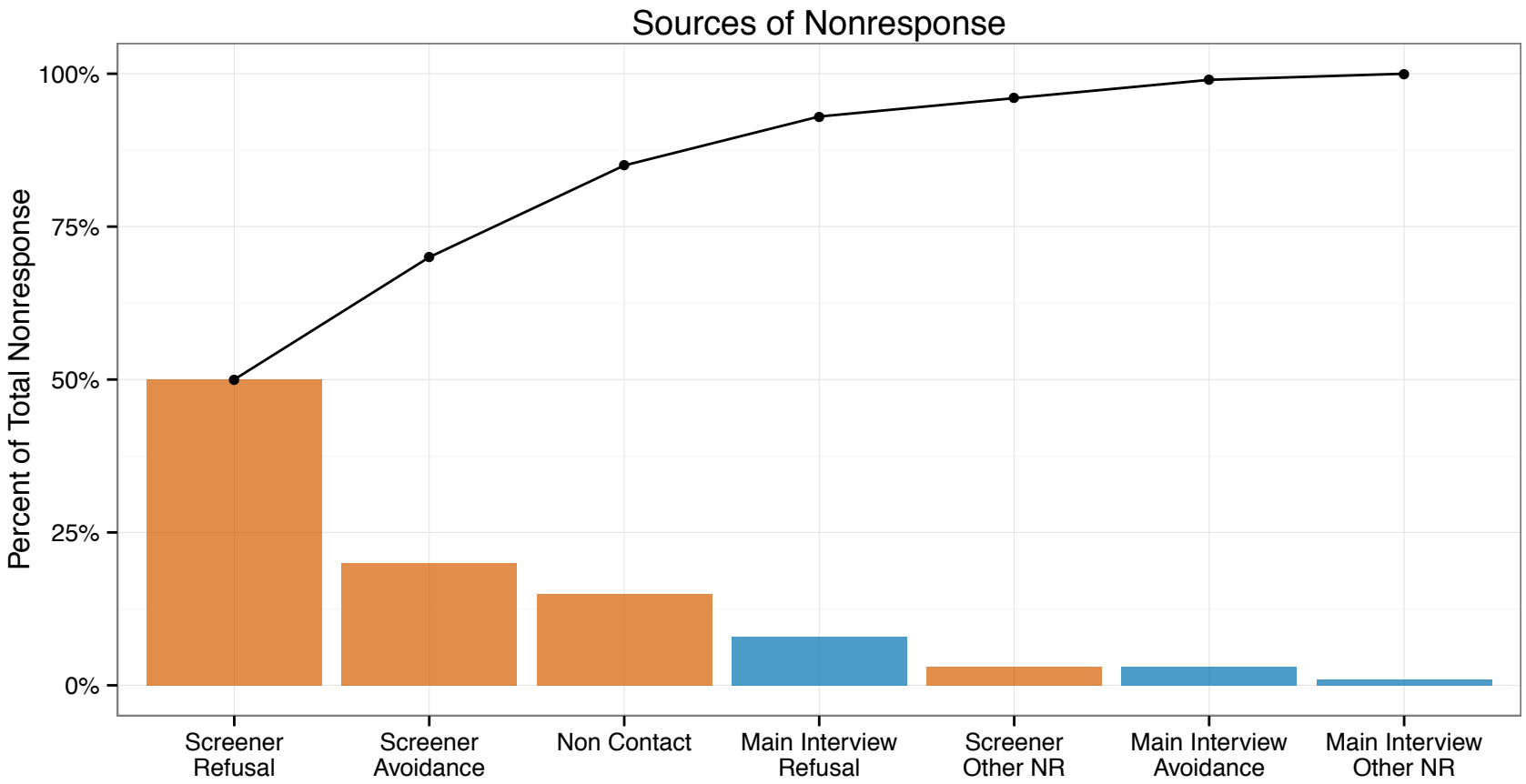
Mercer, 2012

What-If Scenarios
Potential Effects on Estimates in MEPS



Mercer, 2012

Pareto Analysis

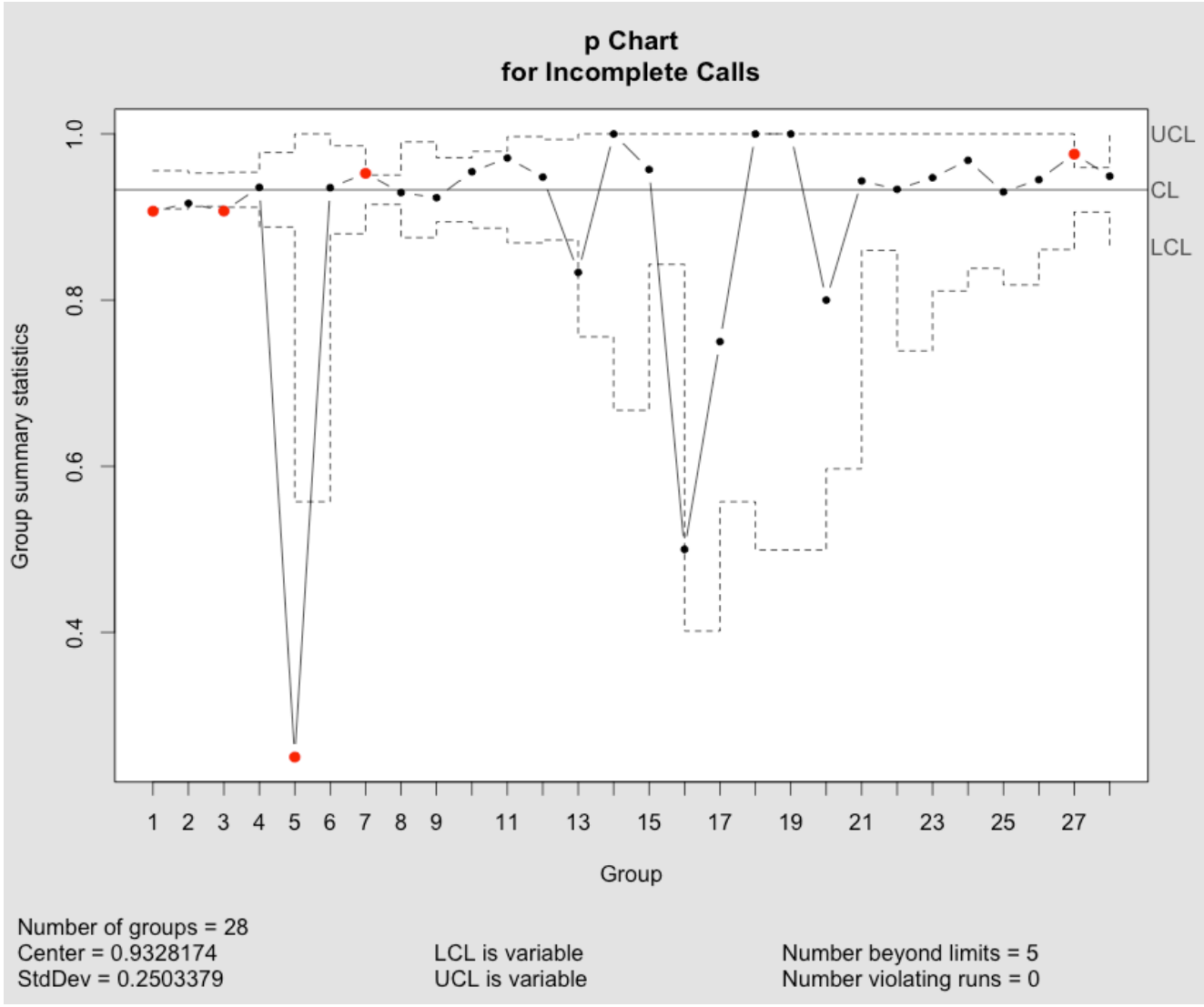


MONITORING DATA COLLECTION

Paradata for Monitoring

- Statistical Process Control
  - Monitor key aspects of data collection process for continuous quality improvement
- Monitoring representativeness in an adaptive design
  - Identify problems
  - Trigger interventions
- Can be incorporated into dashboards

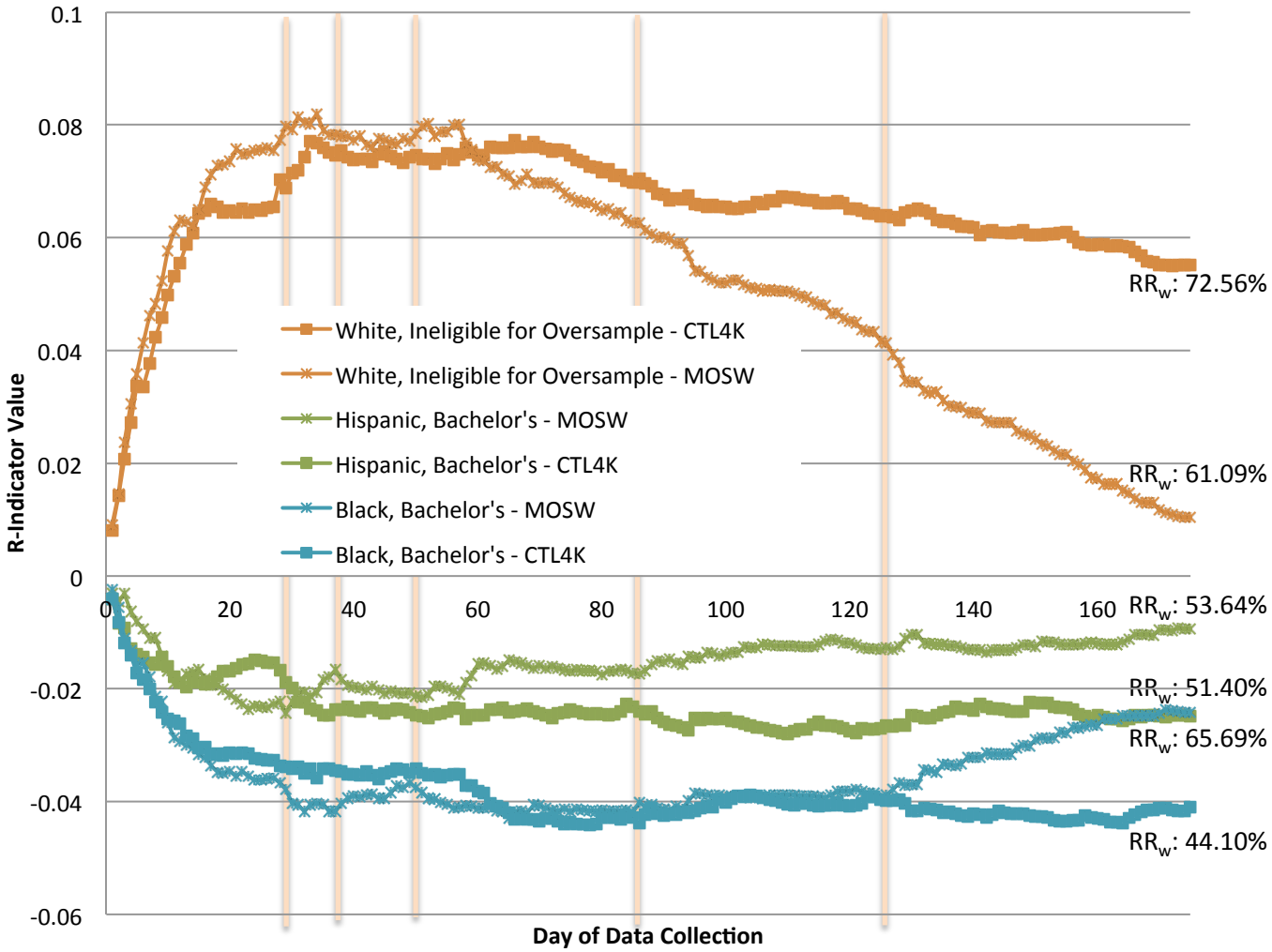
Statistical Process Control



Valliant et al., 2013

Monitoring Representativeness

Unconditional Partial R Indicators for Targeted Subgroups
(Data Through 8/17) Mode Switching vs. Control



Source: National Survey of College Graduates, 2013

Take Home

- There are lots of ways to use paradata with visualization to improve data collection
  - Responsive design or older paradigms like SPC
- Many other aspects of visualization we haven't touched on
  - Interactivity, animation, mapping
- Most important is clarity about what question you want to answer and what data is required to answer it.

Further Reading

Few, S. 2009. *Now you see it: Simple visualization techniques for quantitative analysis*. Oakland, Calif: Analytics Press.

Cleveland, W. S. 1985. *The elements of graphing data*. Monterey, Calif: Wadsworth Advanced Books and Software.

Jans, M., Sirkis, R. and Morgan, D. 2013. "Managing Data Quality Indicators with Paradata Based Statistical Quality Control Tools: The Keys to Survey Performance" in *Improving surveys with paradata: Analytic uses of process information*, ed. F. Kreuter, Hoboken,NJ: Wiley.

References

Couper, M. P. 1998. “Measuring Survey Quality in a CASIC Environment.” In JSM Proceedings, Survey Research Methods Section.

Kreuter, F., Muller, G. and Trappmann, M. 2010. Nonresponse and Measurement Error in Employment Research: Making Use of Administrative Data. Public Opinion Quarterly 74(5): pp. 880-906.

Mercer, A. 2012. Using Paradata to Understand Effort and Attrition in a Panel Survey. In JSM Proceedings, Survey Research Methods Section.

Valliant, R., Dever, J., and Kreuter, F. 2013. Practical Tools for Sampling and Weighting Sample Surveys. New York, NY: Springer.