



# Anticipatory Survey Design: Reduction of Nonresponse Bias through Bias Prediction Models

Andy Peytchev<sup>1</sup>, Sarah Riley<sup>2</sup>, Jeff Rosen<sup>1</sup>,  
Joe Murphy<sup>1</sup>, Mark Lindblad<sup>2</sup>, Paul Biemer<sup>1,2</sup>

<sup>1</sup> RTI International

<sup>2</sup> University of North Carolina

# Acknowledgements

- The work was supported in part by a grant from the National Institutes of Health (Grant R21 HD063070-01A1)
- Appreciation is given to:
  - The Ford Foundation, the University of North Carolina Center for Community Capital, the Community Advantage Program, and RTI's Internal Research and Development Program
  - Community Advantage Program Survey project team at RTI: Barbara Bibb, Brian Burke, Kathleen Considine, Brian Evans, Laura Flicker, and Dawn Thomas-Banks

# Outline

- Need to address nonresponse during data collection
- Previously proposed approach
- New approach
- Modeling (data collection outcomes are forthcoming)

## Directing Data Collection Effort

- Maximize response rates – whether used as an indicator of nonresponse bias or of the general notion of representativeness
- Increased effort motivated by desire to increase or maintain response rates:
  - Targeting of “easier” sample cases
    - As a survey management decision
    - As interviewers strive to achieve response rate goals
  - Continued use of the original protocol
- Such an approach can fail to reduce nonresponse bias
  - “More of the same”

# Underlying Weakness of Strategy

- Implicit orthogonality

$$E[\bar{y}_r - \bar{y}_s] = E\left[\frac{m_s}{n_s}(\bar{y}_r - \bar{y}_m)\right]$$

- Stochastic view

$$\text{Bias}(\bar{y}_r) \approx \frac{\sigma_{y,\rho}}{\bar{\rho}}$$

Where  $\rho$ -bar is also the response rate

- Increasing  $\rho$ -bar by targeting high propensity cases may be even less successful in decreasing bias, by increasing  $\text{Cov}(y,\rho)$

# Intervention on Sample Members with Low Predicted Response Propensity (Peytchev, Riley, Rosen, Murphy, and Lindblad, 2010)

Decrease through (low)  $\rho$ 's:

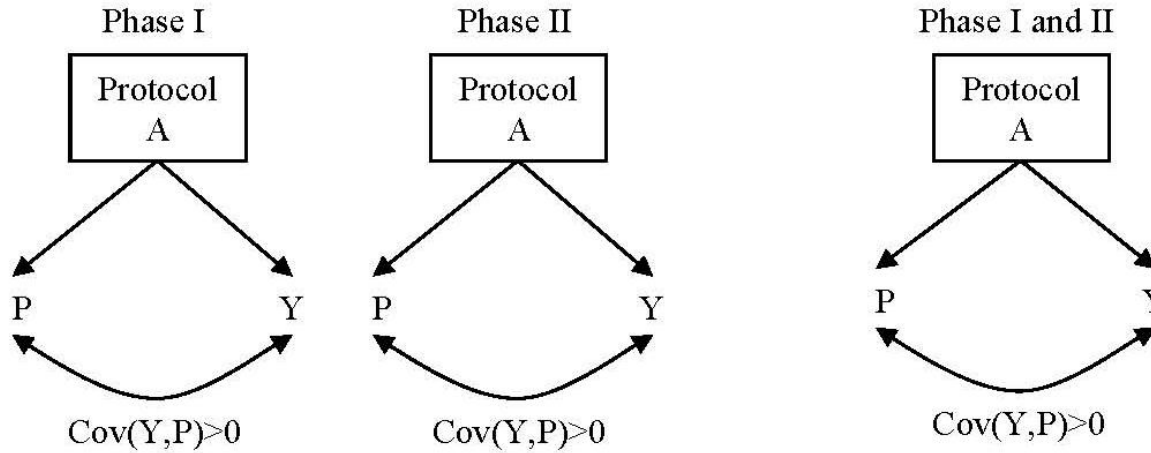
$$\sigma_{y,\rho} = E[\text{Corr}(y, \rho) \sqrt{\text{Var}(y) \text{Var}(\rho)}]$$

Can remain the same

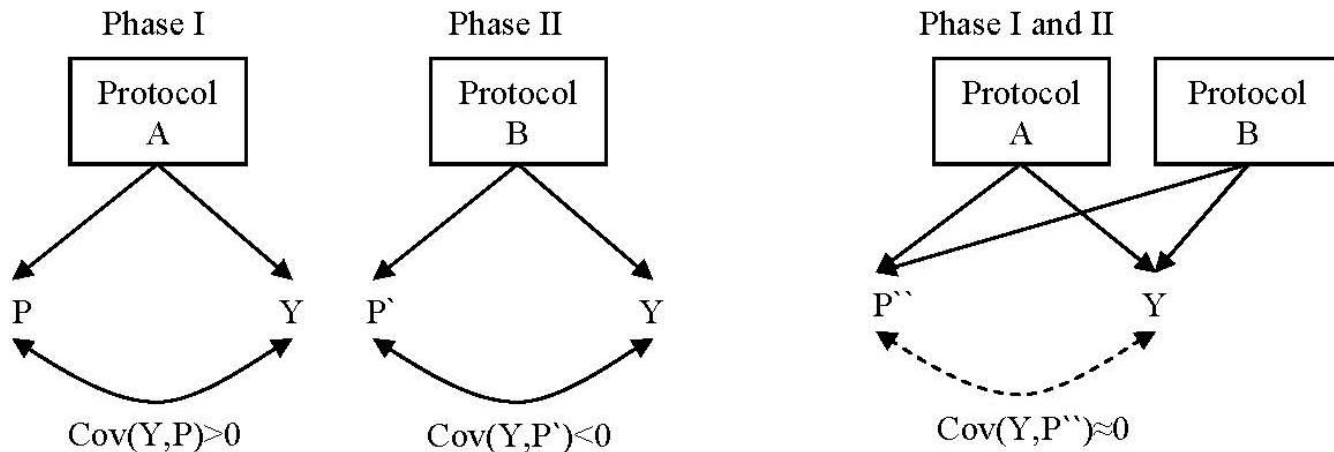
$$\text{Bias}(\bar{y}_r) \approx \frac{\sigma_{y,\rho}}{\bar{\rho}}$$

# Implementation of Approach in Multiphase Design for Nonresponse Bias Reduction

## A. Same Protocol



## B. Proposed Approach



# Necessary Components in the Proposed Approach

## 1. Estimation of $\rho$

- Frame information
- Paradata during current data collection
- Prior waves of data collection
  - Demographic
  - Substantive variables (i.e.,  $y_{t-1}$ )
  - Paradata

## 2. Intervention on cases with low $\hat{\rho}$

- Informed by empirical findings
- Informed by embedded experiments in the current study with option to stop either control or experimental conditions (a feature of responsive design)

## Experimental Evaluation in CAPS

- We tested this approach in Wave 5 of the Community Advantage Panel Survey, which evaluates this secondary mortgage program.
- Data collection was conducted between July 2008 and January 2009.
- Longitudinal panel study design:
  - Two samples (owners and renters)
  - Two modes of data collection (in-person and telephone). Only the in-person sample is used here.

# The Prioritization and Intervention Steps

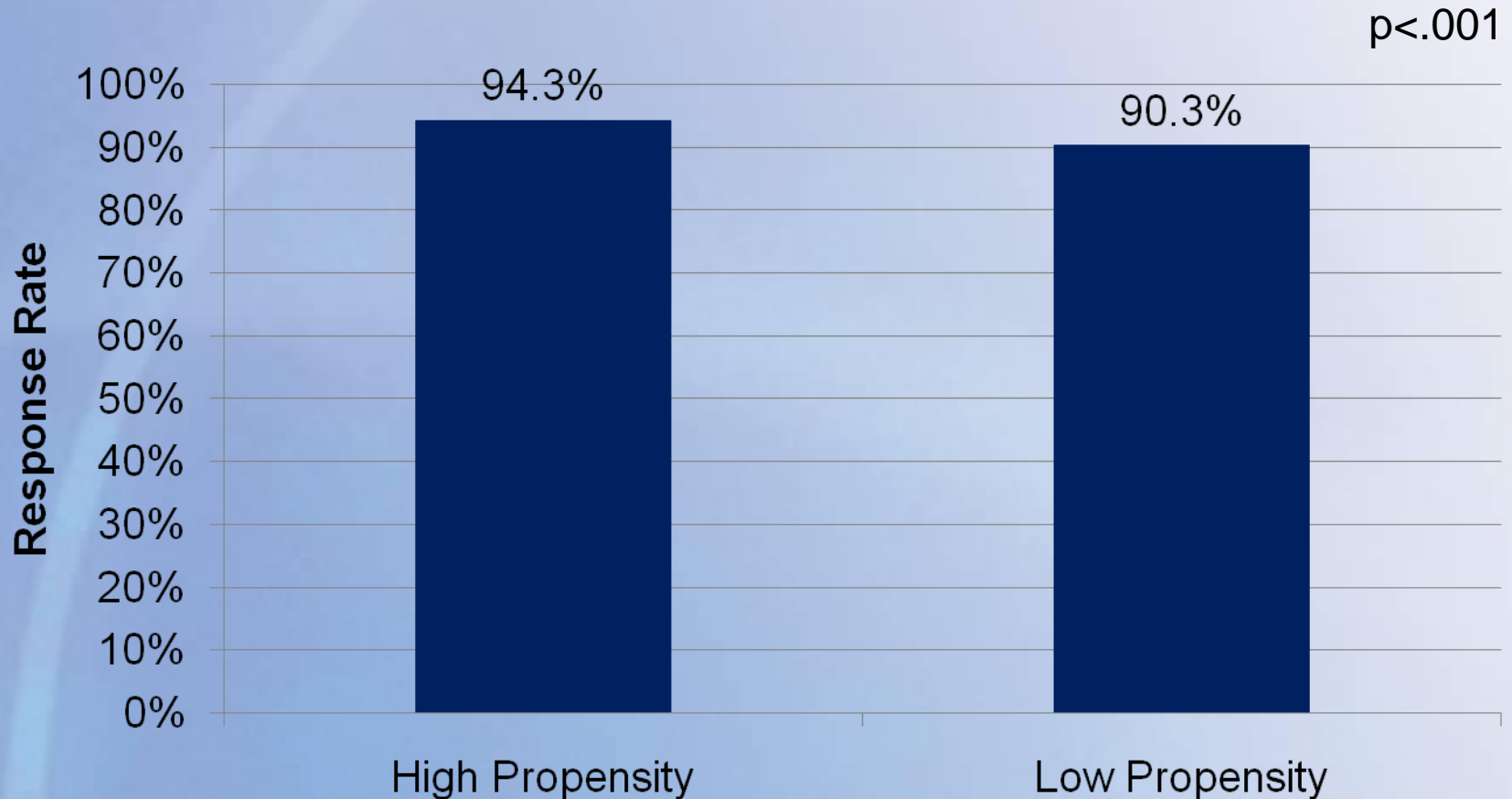
## 1. Estimation of $\hat{\rho}$ and prioritization

- Response propensities were estimated and sample was classified into low and high response propensity, with:
- Demographic characteristics: age, race, education, gender
- Key variables at baseline: mortgage delinquency, time since loan origination, and time since purchasing the home
- Voting: whether respondent had voted in the 2000 election
- Prior wave paradata: whether respondent reported not being interested, and whether respondent had ever hung up during the introduction

## 2. Intervention on cases with low $\hat{\rho}$

- Low propensity sample cases were randomly assigned to control and priority groups, within geographic area
- Additional interviewer incentive of \$10 per completed interview

# 1. Evaluating $p^{\wedge}$ : Proportion Interviewed by (Predicted) Propensity Group



Difference was as high as 10 percentage points during data collection (end of Phase 1).

## 2. Evaluating the Approach: Intervention on the Response Propensities

- Response rate ( $p=.56$ )
  $RR_{Control}=89.8\%$   
 $RR_{Exper.}=90.8\%$
- Mean propensity ( $p=.33$ )
  $Mean(\rho_{Control})=.915$   
 $Mean(\rho_{Exper.})=.917$
- Variation in propensities ( $p<.001$ )
  $Var(\rho_{Control})=.00484$   
 $Var(\rho_{Exper.})=.00292$
- No difference in means and variances of propensities estimated prior to data collection
- **No reduction in estimated  $Cov(y,\rho)$  for key survey variables**
- **No reduction in estimated nonresponse bias**

# Interim Summary

So far:

- Propensity scores predictive of likelihood of completed interview (despite assignment within interviewer)
- Incentive-based interviewer intervention seems ineffective, as tested here
  - Same effort
  - Same response rates
  - Reduced variability of propensities, but likely spurious, as it did not (1) reduce  $Cov(y,\rho)$  and (2) reduce estimated nonresponse bias

Next steps:

- Can alter prioritization strategy to target bias more directly
- Can implement a respondent-based intervention

## New Approach

- Targeting all cases with low  $\hat{\rho}$  (i.e., *independent of y*) can be inefficient in reducing *nonresponse bias* (yet, depending on the adjustment model, very efficient at reducing *nonresponse variance* in adjusted estimates)
- **Prioritize cases based on predicted  $y$ , for values of  $y$  that are associated with low  $\hat{\rho}$  (i.e., underrepresented among respondents)**
- If the prediction models perform well and the intervention on these cases is effective, this should work to directly **minimize  $Cov(y,\rho)$**

Recall that: 
$$Bias(\bar{y}_r) \approx \frac{\sigma_{y,\rho}}{\bar{\rho}} !$$

## New Approach Continued

- Bias, for example, in the expected value of a proportion, is zero when the response rates are equal for  $y=0$  and for  $y=1$ . This approach aims to accomplish this goal through models estimating  $y$  and  $\rho$ . As shown in Biemer, 2008:

$$Bias(\bar{y}_r) = \frac{(\bar{\rho}_{y=1} - \bar{\rho}_{y=0})P(1-P)}{\bar{\rho}} \quad \text{thus, } Bias(\bar{y}_r) = 0$$

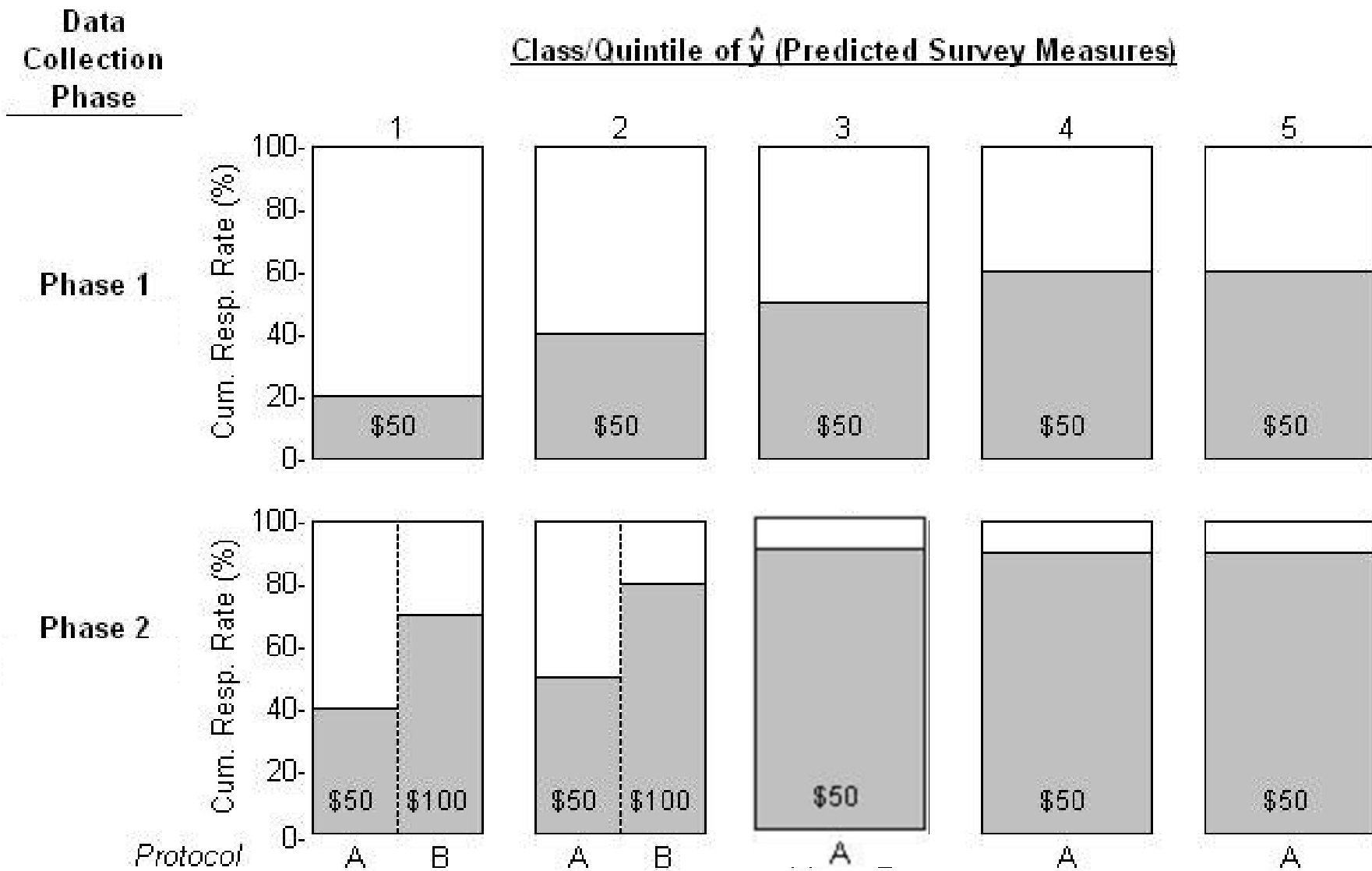
when  $\bar{\rho}_{y=1} = \bar{\rho}_{y=0}$

- To the extent that nonresponse adjustment models include variables associated with both  $\rho$  and  $y$ , nonresponse variance due to postsurvey adjustments should also be minimized, compared to reliance on adjustments alone.

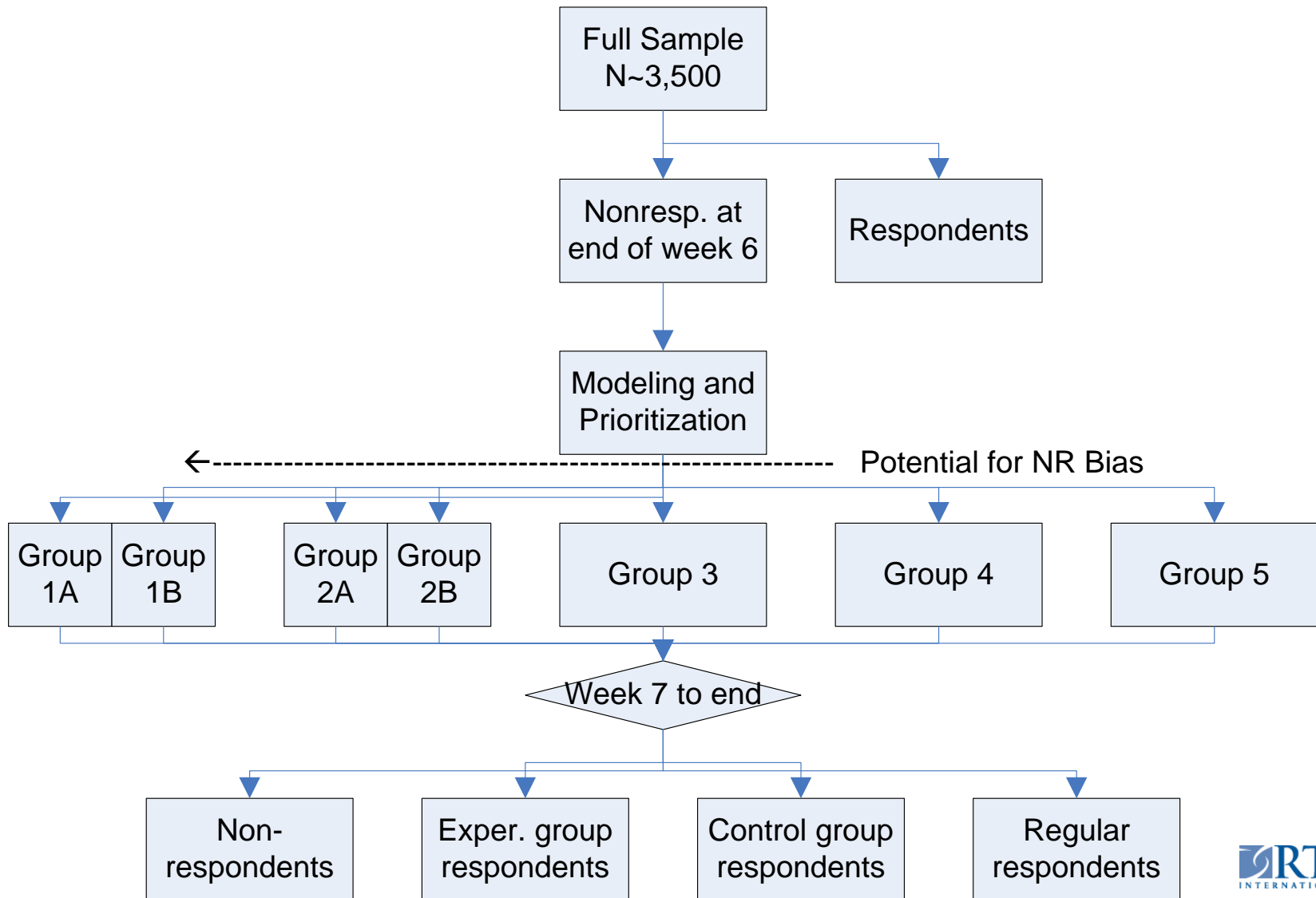
## Prioritization Models

- Estimate predicted response propensities for all cases
- Conduct a factor analysis using the key survey variables (this is a data reduction step to allow for multiple key survey variables)
- Regress the main factor on the predicted response propensity (this step aims to maximize the ability to differentiate cases with different predicted responses, based on the likelihood to respond)
- Create quintiles based on the predicted factor scores
- Within the quintiles with the lowest mean response propensity, randomly assign cases to experimental and control groups

# Schematic of Design & Expected Outcomes



# Data Collection Design in the 2011 Experiment



## Progress So Far...

- Models fit reasonably well
  - Response propensities based on current outcomes during data collection were quite different in one of the samples
- Telephone data collection will likely result in greater nonresponse than in the first experiment, allowing for greater room for improvement in nonresponse bias
- Currently in Phase 1 of data collection
- Expect to have initial results towards the end of the year

# Contact

Andy Peytchev: [apeytchev@rti.org](mailto:apeytchev@rti.org)