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

Andy Peytchev[1], Sarah Riley[2], Jeff Rosen[1], Joe Murphy[1], Mark Lindblad[2], Paul Biemer[1,2]

[1] RTI International
[2] 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

- Preliminary results

- Future research

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 (cost)
    - 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"

Intervention on Sample Members with Low Predicted Response Propensity
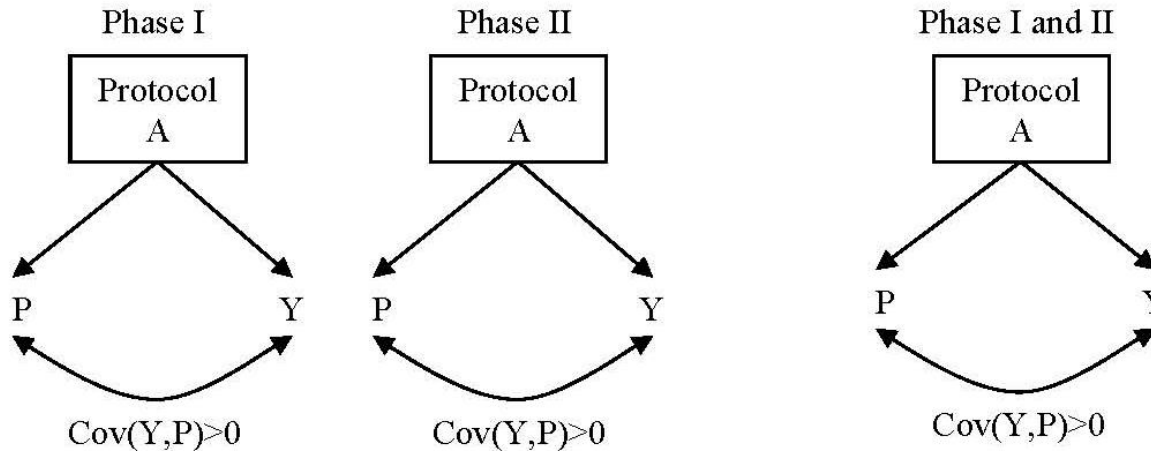
Decrease
through (low) ρ's:

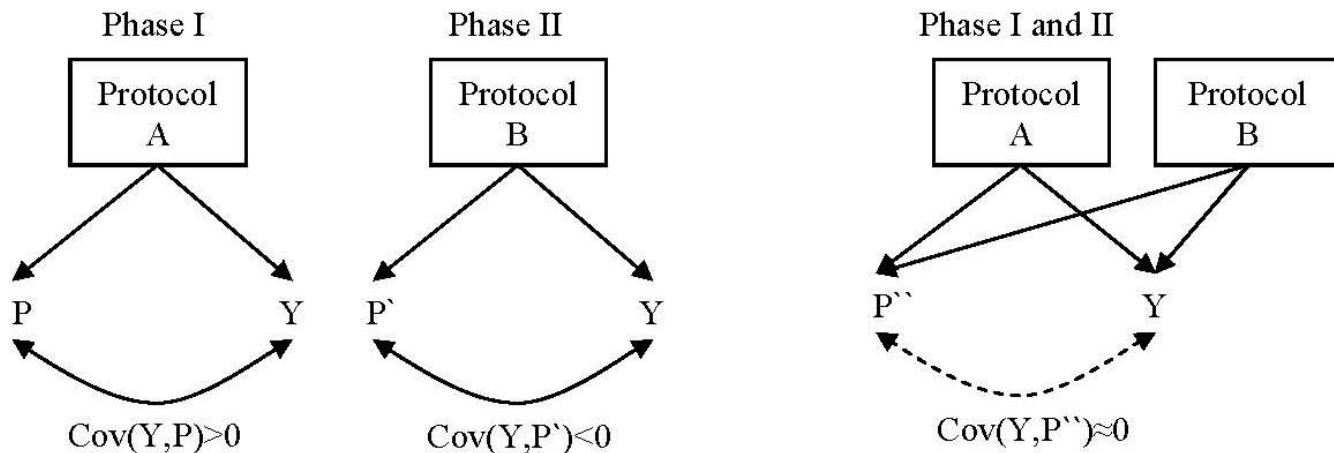$$\sigma_{y,\rho} = E[Corr(y,\rho)\sqrt{Var(y)Var(\rho)}]$$

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

Can remain
the same

Implementation of Approach in Multiphase Design for Nonresponse Bias Reduction

A. Same Protocol

Phase I — Protocol A → P, Y; $Cov(Y,P) > 0$

Phase II — Protocol A → P, Y; $Cov(Y,P) > 0$

Phase I and II — Protocol A → P, Y; $Cov(Y,P) > 0$

B. Proposed Approach

Phase I — Protocol A → P, Y; $Cov(Y,P) > 0$

Phase II — Protocol B → P`, Y; $Cov(Y,P`) < 0$

Phase I and II — Protocol A, Protocol B → P``, Y; $Cov(Y,P``) \approx 0$

Necessary Components in the Proposed Approach

1. Estimation of ρ
   - 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

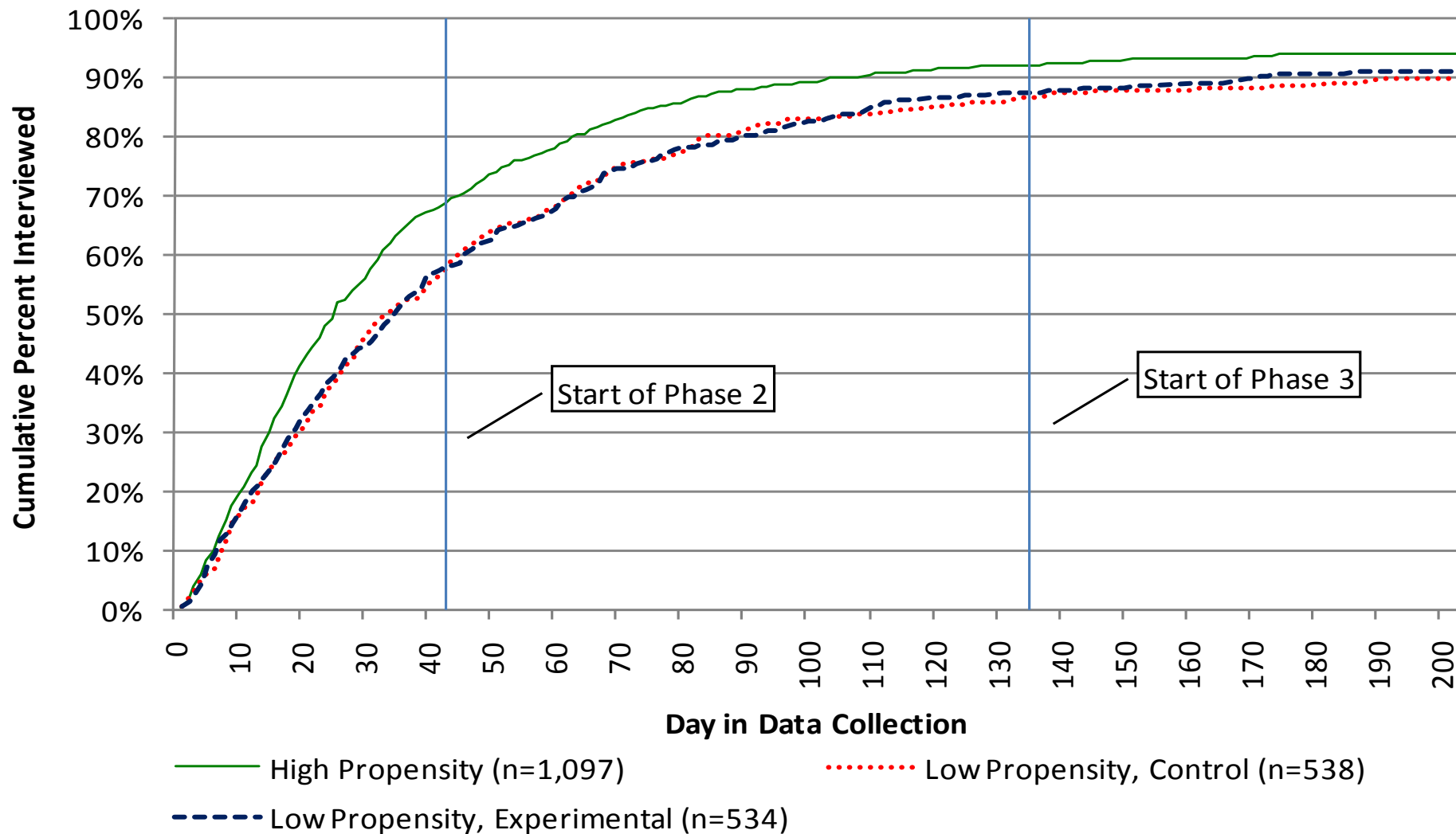# 1. Evaluating ρ^: Proportion Interviewed by (Predicted) Propensity Group
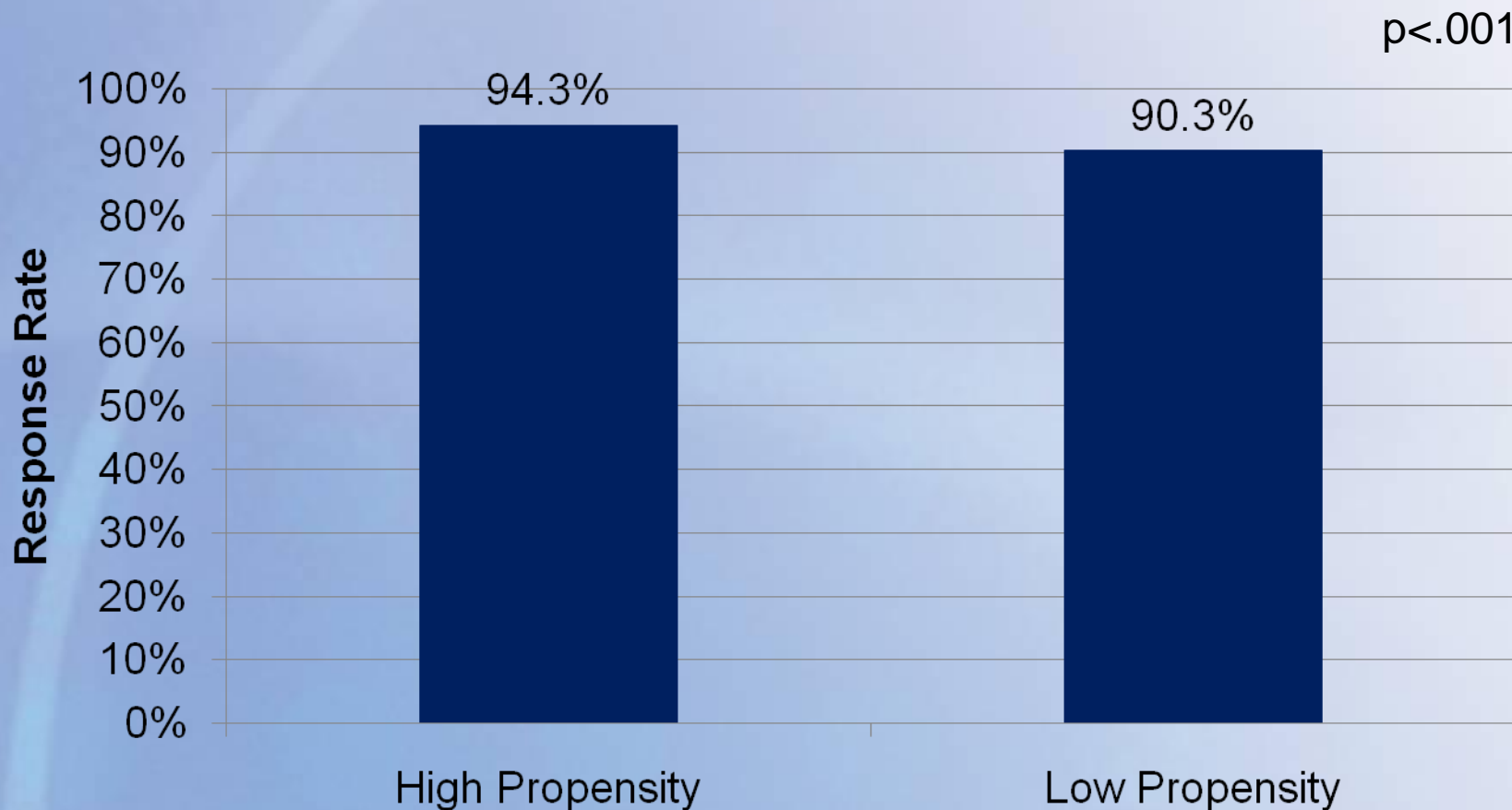
Difference was as high as 10 percentage points during data collection (end of Phase 1).

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,ρ) for key survey variables
- No reduction in estimated nonresponse bias

RTI
INTERNATIONAL

Cumulative Number of Interviews Completed by Day

Chart: Cumulative Percent Interviewed (y-axis, 0%–100%) vs. Day in Data Collection (x-axis, 0–200). Vertical lines mark "Start of Phase 2" (~day 43) and "Start of Phase 3" (~day 135).

Legend:
— High Propensity (n=1,097)
······· Low Propensity, Control (n=538)
--- Low Propensity, Experimental (n=534)

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:

- Alter prioritization strategy to target bias more directly

- 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,ρ)*

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

Intervention Revisited

Decrease
through (low) ρ's:

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

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

Can remain
the same

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 *ρ*. As shown in Biemer, 2008:

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

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 *ρ* 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 and Objective
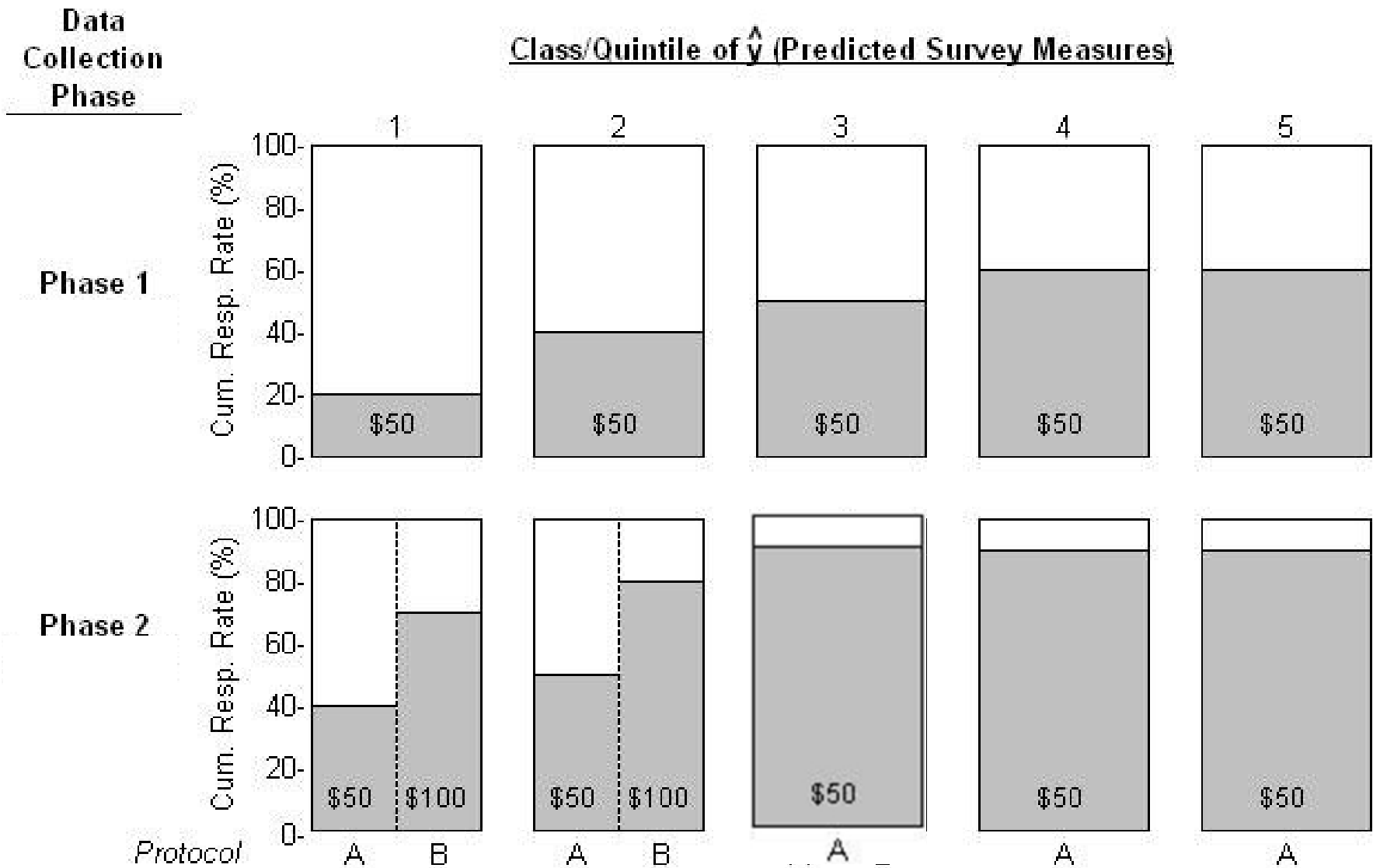
Data and Methods

- Community Advantage Panel Survey, 2011 (wave 8)
- Two samples, home owners and renters
- All interviews conducted by telephone in this wave (with lower expected response rate)

- Start data collection
- Estimate prioritization models
  - Models for y
  - Models for ρ
- After a certain period, implement higher respondent incentives for nonresponding cases that are likely contributing to nonresponse bias, based on models
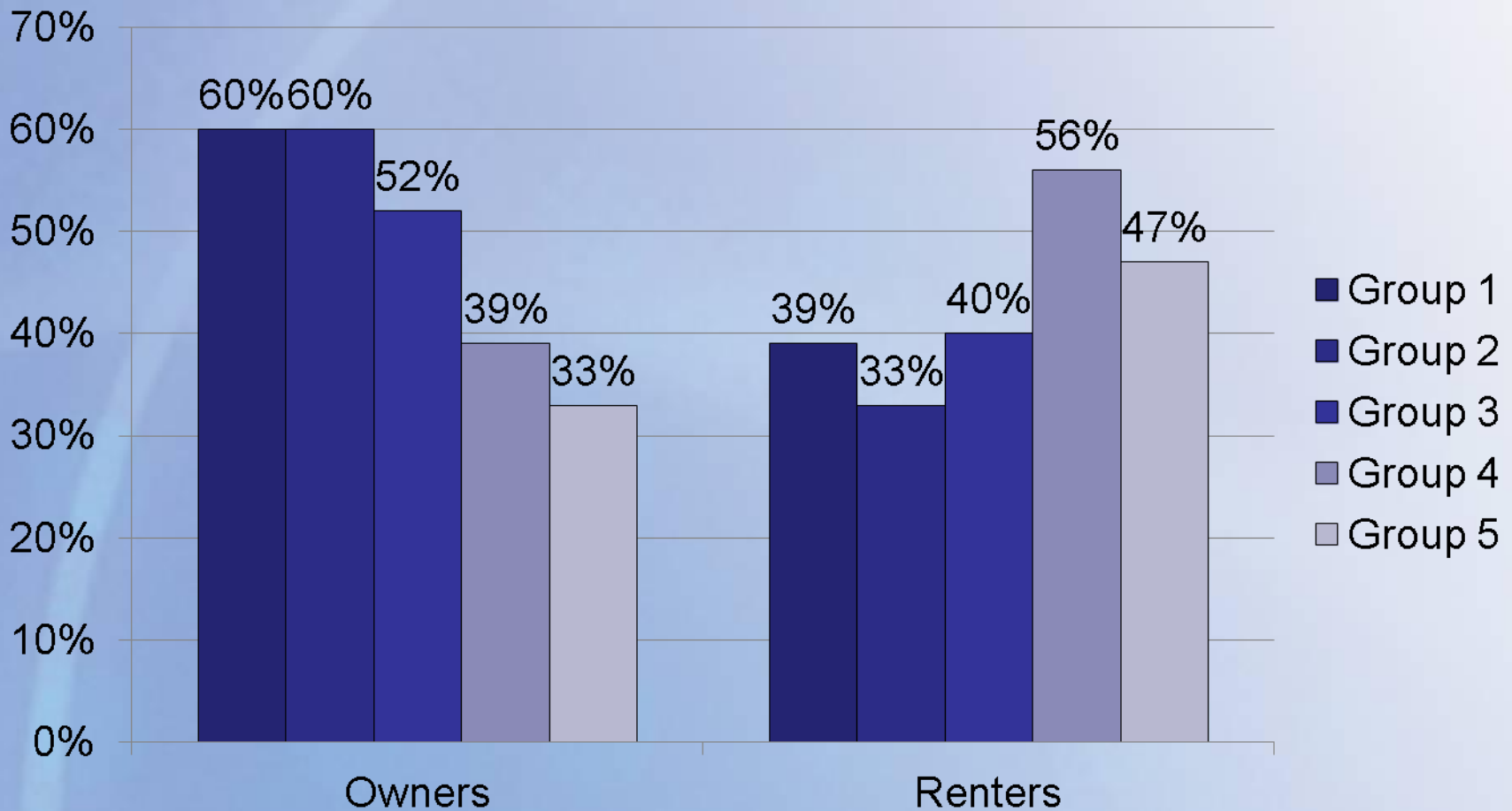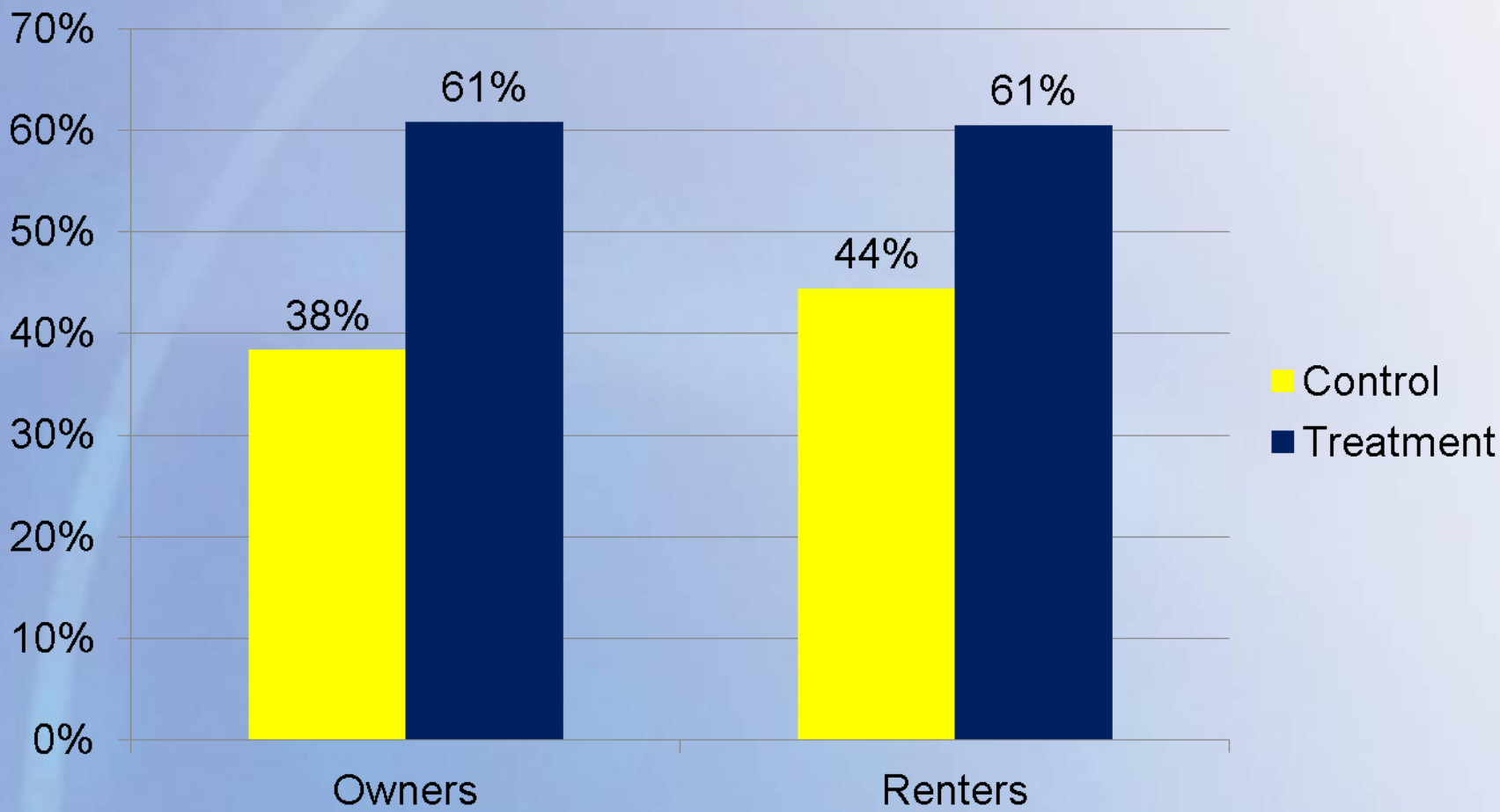
Data Collection Design in the 2011 Experiment
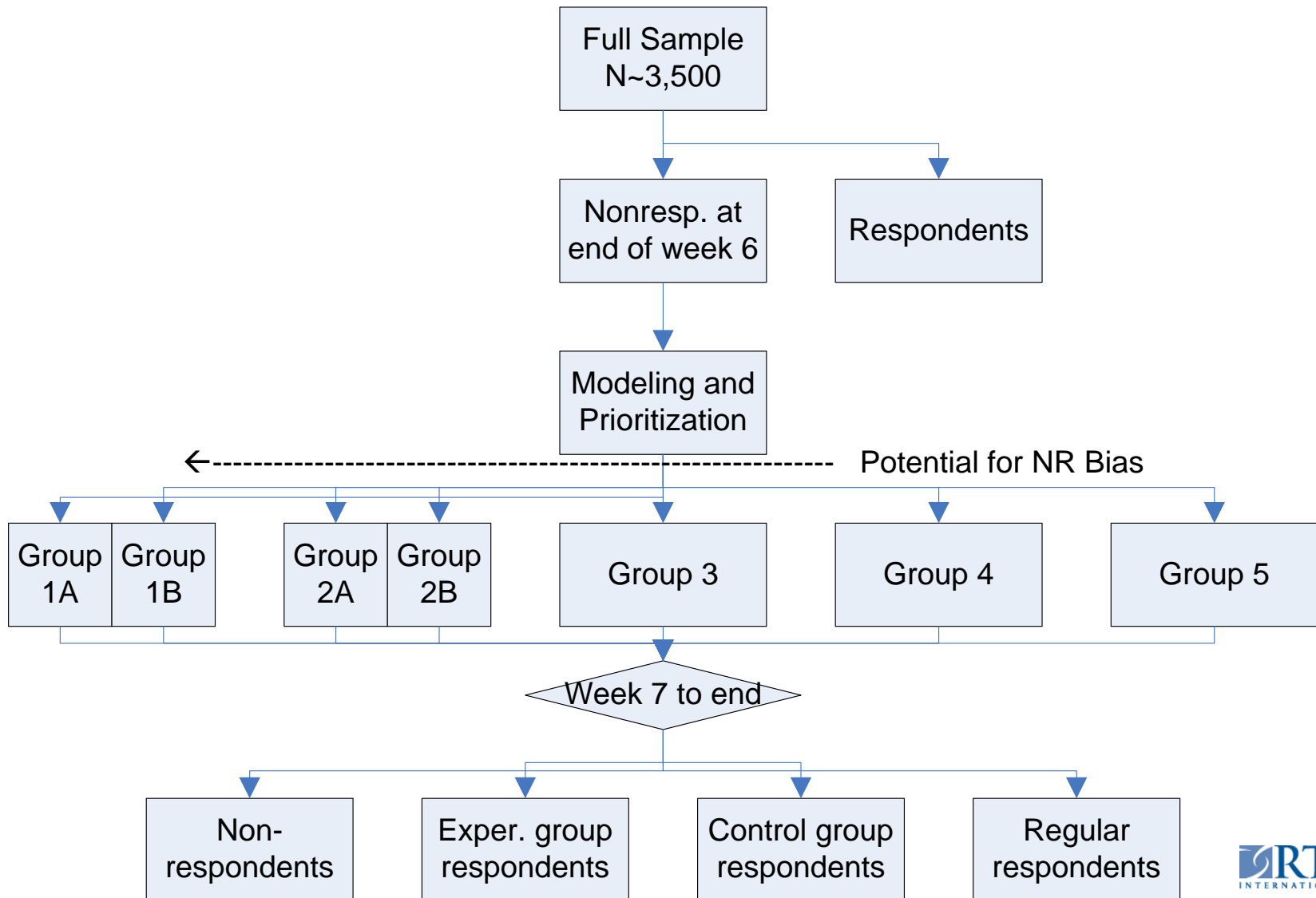
Progress So Far… October 25

- Models fit reasonably well
  - Response propensities based on current outcomes during data collection were quite different in one of the samples
    - Current outcomes models may be more timely (same wave of data collection), but yield propensities with very different meaning early in data collection
  - Relationship between principal component scores and response propensities was different (opposite in direction) in the two samples –- y-ρ relationship is different for subpopulations

RTI
INTERNATIONAL

Interview Rates among Nonexperimental Cases

Owners:
- Group 1: 60%
- Group 2: 60%
- Group 3: 52%
- Group 4: 39%
- Group 5: 33%

Renters:
- Group 1: 39%
- Group 2: 33%
- Group 3: 40%
- Group 4: 56%
- Group 5: 47%

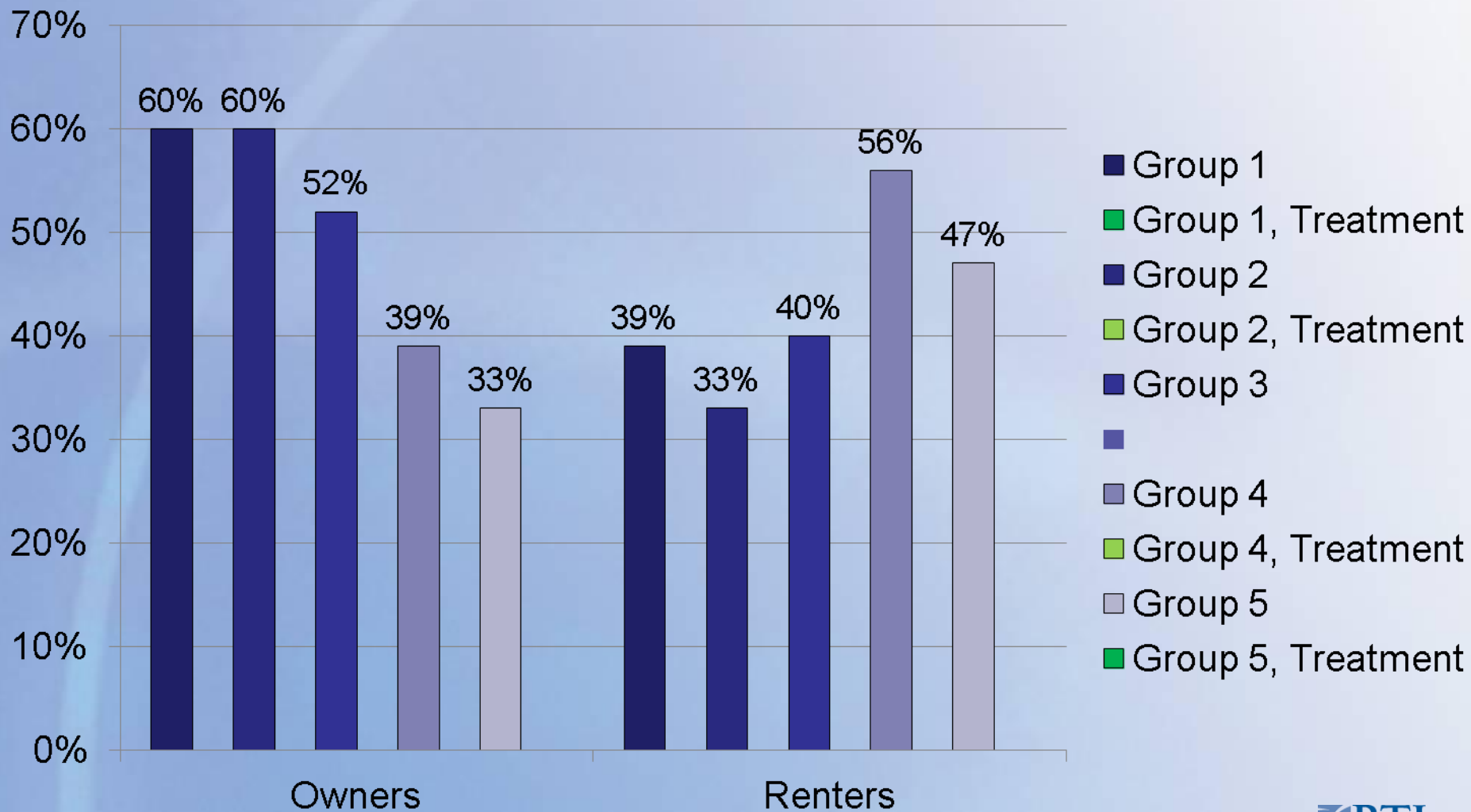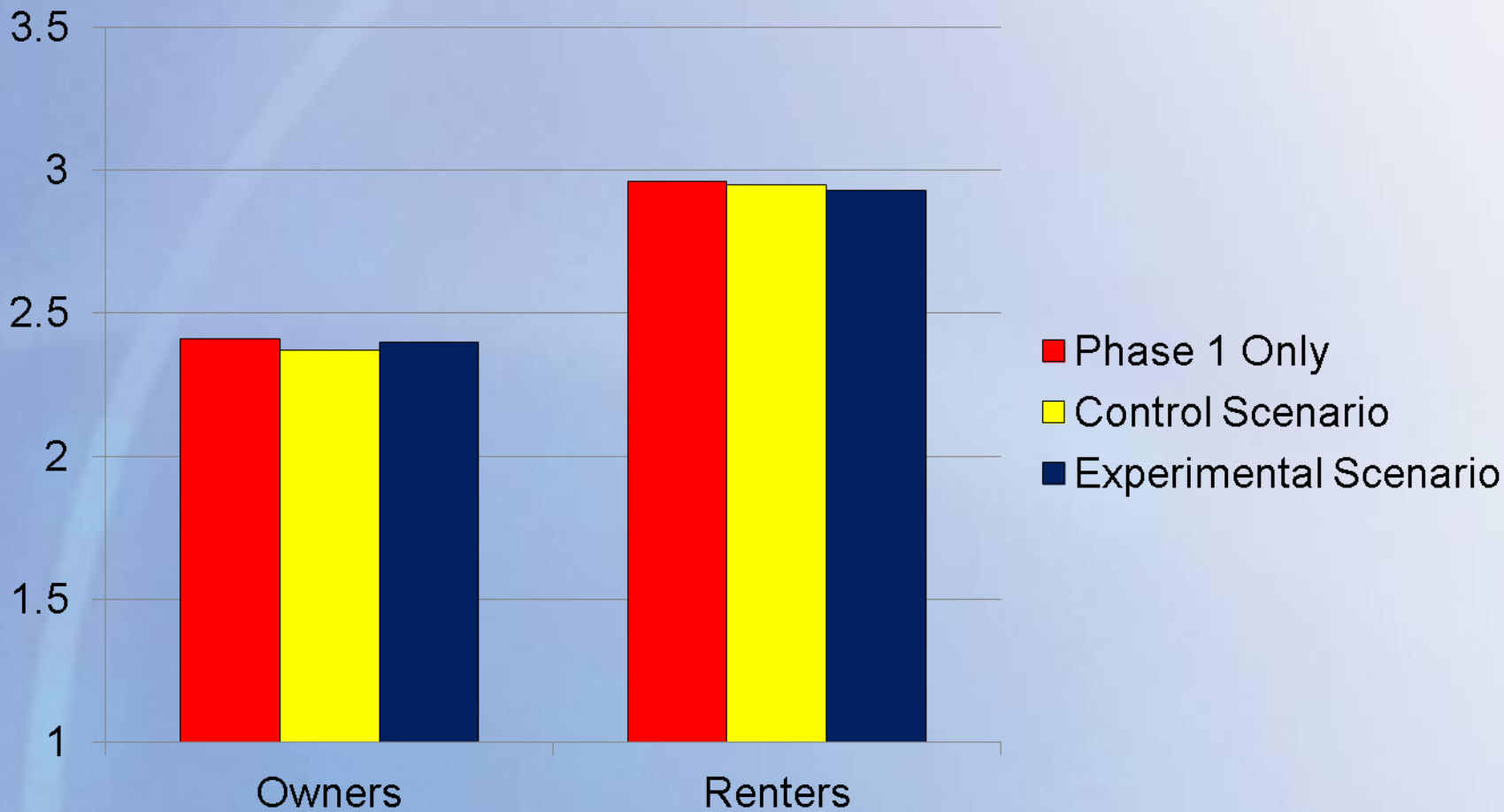# Interviews among Experimental Strata, by Experimental Condition
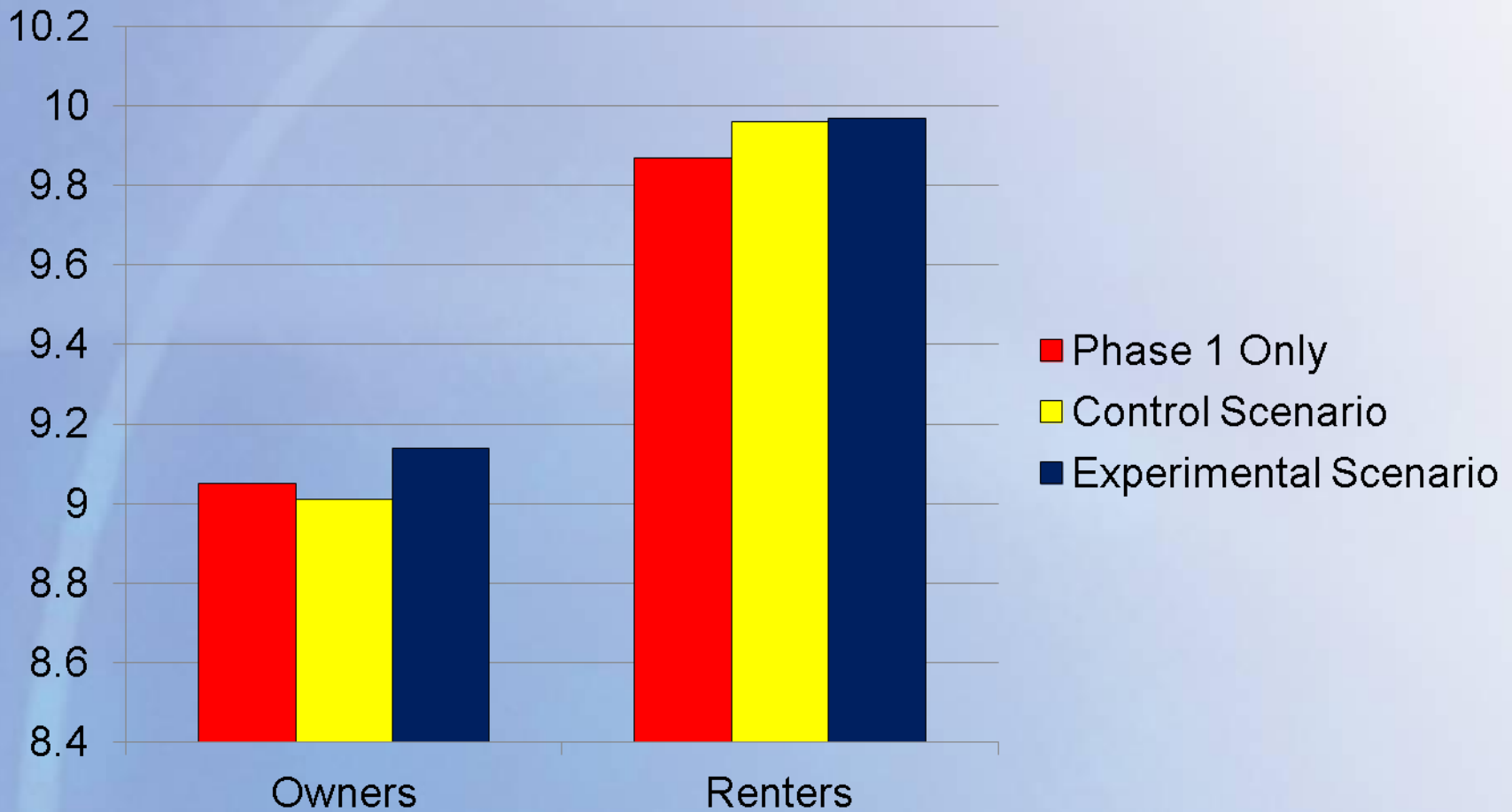
Interview Rates among All Cases
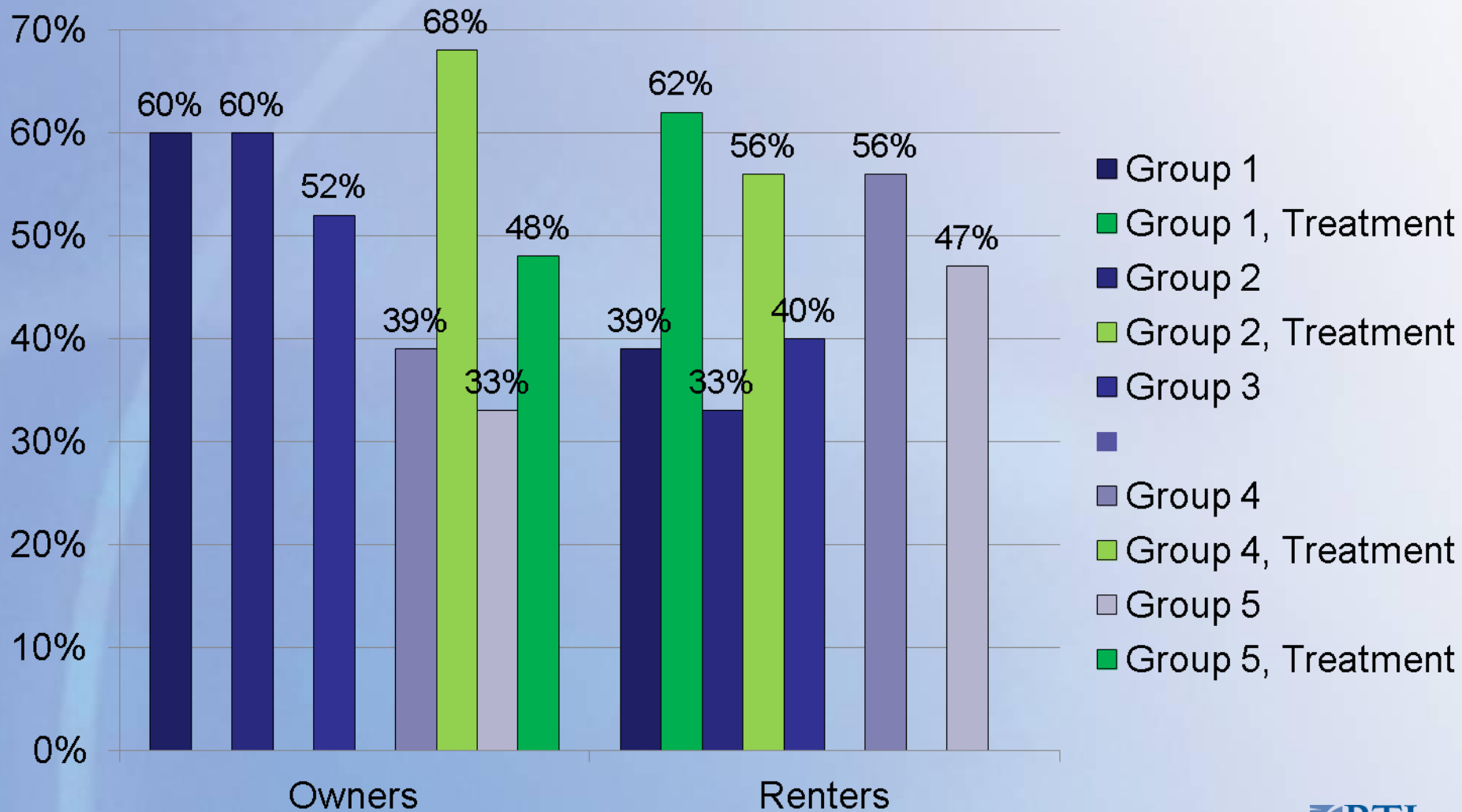
Proportion Working
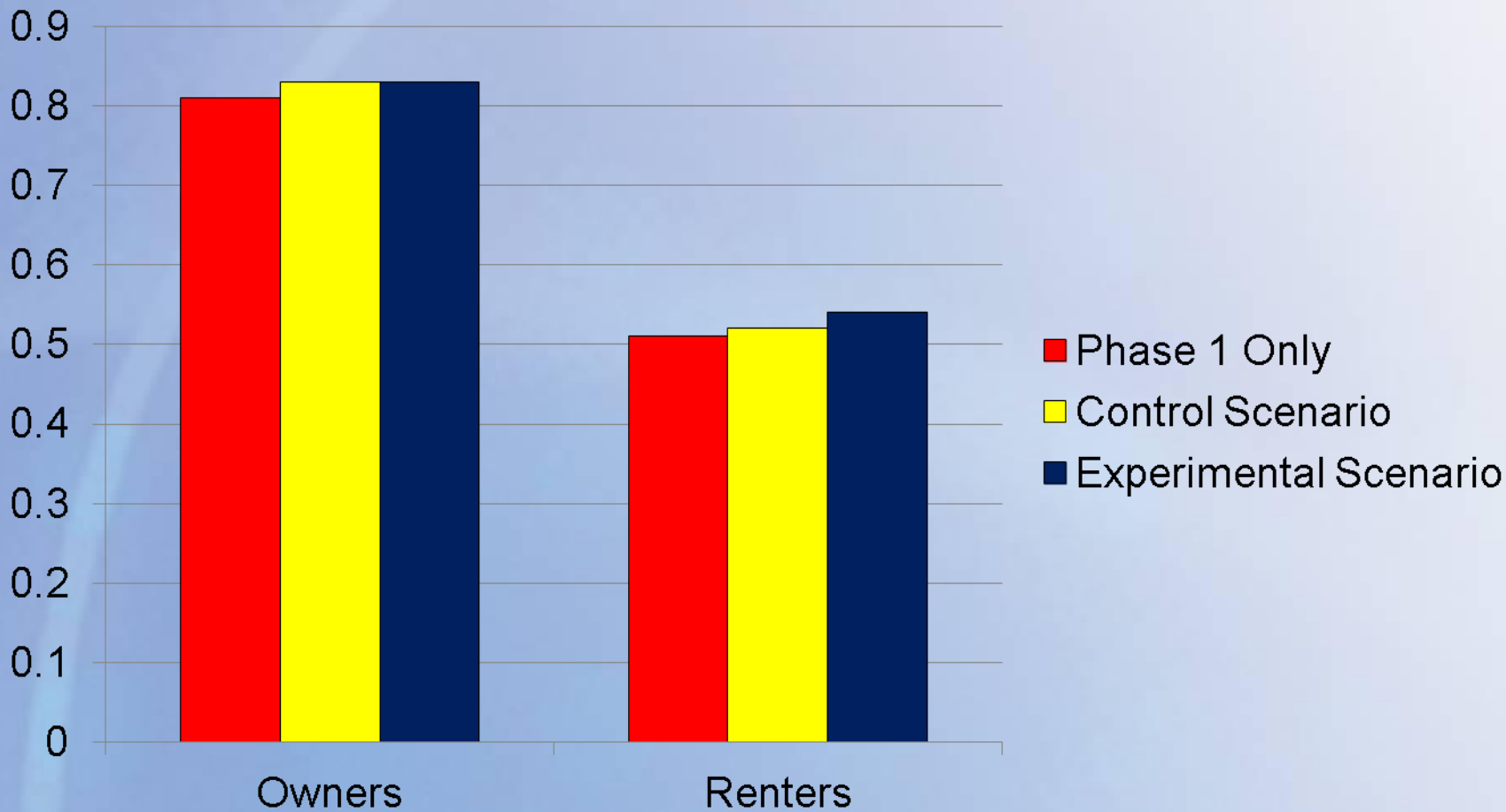
General Health

Stress Scale

Summary

- With the panel survey design, we were able to estimate response propensities that were quite predictive of the outcome in the following wave.

  - Both interim response propensities (allow better use of paradata and responsiveness during data collection) and propensities for outcomes in the prior wave can be estimated (allow for intervention to start at the beginning of data collection).

- The interviewer incentivisation did not prove to be effective (in a face to face administration), but respondent incentives were effective (in a telephone administration).

Summary - continued

- Because the interviewer incentive was not effective, we were not able to establish whether pursuit of low propensity cases can reduce bias.

- A similar but more direct approach to bias reduction is to target cases with low propensities that are likely to be different on key survey measures

- The respondent incentive has been successful so far in equalizing response rates across groups defined by key survey variables

- No substantial difference in estimates between control and experimental conditions
    - Three months remaining
    - Need for predictors and explicit modeling of change

Future Directions for Research

- Which response propensities to use? Use both? During data collection or those using outcomes in prior waves?

  - May depend on when the different protocol is implemented – yet this is a continuum in time during data collection. Also depends on the amount and properties of additional data that can be collected. Ultimately, it is also a theoretical question.

- Target cases based on response propensities or key survey variables with values associated with lower propensities? Low response propensities alone are

  - Easier to implement, but
  - May not be direct enough for reduction of nonresponse bias.
  - May also depend on the objectives – bias vs. variance reduction.

RTI
INTERNATIONAL

Future Research - continued

- Approach relies heavily on auxiliary data informative of likelihood of participation, and key survey variables
  - Less challenging for surveys with a panel design, but then the need for correlates of change are needed. This area has not received sufficient attention.

- Particularly for cross sectional survey designs, different models and interventions may need to be set up for noncontacts and for refusals.

- Continuous experimentation with models and interventions is essential to optimizing such responsive designs for reduction of nonresponse bias and variance, rather than fixing the design.

RTI
INTERNATIONAL

Contact

Andy Peytchev: apeytchev@rti.org