Simulating outcomes: a Bayesian approach to estimating longitudinal survey parameters

Michael A. Duprey, Daniel J. Pratt, Darryl A. Cooney, T. Austin Lacy

BADEN 5th Workshop on Adaptive and Responsive Survey Design
November 7, 2017, University of Michigan

Summary

Presentation outline

- Case study background
  - Introduction of High School Longitudinal Study of 2009 (HSLS:09)
  - Calibration subsample
  - Phases
  - Subgroups
  - Modeling
  - Adaptive components
- Simulation case study
  - Goals
  - Approach
  - Priors
  - Posterior results
- Discussion
- Next steps

Case study background

High School Longitudinal Study of 2009 (HSLS:09)

- U.S. Department of Education, National Center for Education Statistics
- Nationally representative, longitudinal study of 23,000+ 9th graders in 2009
- Study design:
  - Base year (2009)
  - First follow-up (2012)
  - 2013 Update (2013)
  - **Second follow-up (2016)**

HSLS:09 second follow-up summary

- **Calibration** subsample and **main** sample
- Several distinct **phases** of data collection with corresponding interventions
- 3 study **subgroups** of interest
- **Response propensity model** to maximize efficient allocation of project resources
- Model to predict likelihood of contributing to **nonresponse bias**, used to target sample members for interventions

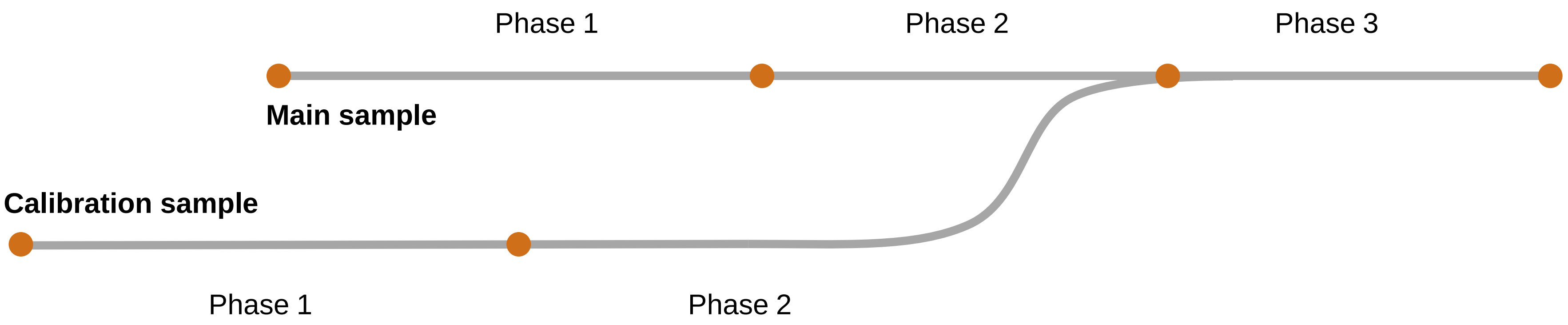
Case study background

Calibration subsample

- Fielded **8 weeks** in advance of the main sample to **experimentally test** the effectiveness of planned interventions in phase 1 and phase 2
- Subsample of about 15 percent of total sample, n = 3,300
- Interventions applied over **3 phases** of interest

Phases

- **Phase 1:** baseline incentive (monetary)
- **Phase 2:** Incentive *boosts* (monetary)
- **Phase 3:** Field interviewing (CAPI), abbreviated interview; calibration and main sample aligned



Case study background

Subgroups

- Stratified based on **previous-round experience** with cohort
- Differentiated so that customized interventions could be developed (informed by prior rounds), and applied to subgroups independently

Second follow-up subgroups

Subgroup A
High school late / alternative / noncompleters

Had not completed HS; were still enrolled in HS; received alt. credential; completed HS late; had dropout with unknown HS completion status

Subgroup B
Ultra-cooperative respondents

HS completers that participated in base year, first follow-up, and 2013 Update without incentive offer

Subgroup C
All other cases

Early / on-time regular diploma completers (not subgroup B) and cases with unknown HS completion status (not subgroup A)

Case study background

Modeling

Response propensity model

Estimates unit-level response probability

- **Covariates:** Model covariates combine paradata (prior-round paradata and demographics) and key variables of interest found to maximize prior-round response prediction
- **Dependent variable:** 2013 Update response (immediately prior round)
- **Estimation:** Once, prior to data collection start

Application

Used in phase 3 (field interviewing) to exclude pursuit of low response propensity cases

Bias likelihood model

Identifies nonrespondents in the most underrepresented groups

- **Covariates:** Chosen such that differences should proxy nonresponse bias; *excludes paradata*
- **Dependent variable:** Current-round response
- **Estimation:** Re-estimated throughout data collection, before intervention deployment

Application

Used in phase 2 (boosts) and phase 3 (field interviewing) to target cases for incentive increases and field interviewing

Case study background

Calibration-informed adaptive components

Incentives selected

- Best-performing incentives (baseline and boosts) were offered to main sample cases

Redefined subgroup B

- To conserve project resources, the subgroup B (ultra-cooperative) set definition **was expanded** for the **main study** only
- Expanded definition sought to identify **relatively homogeneous** group of highly cooperative sample members within subgroup C (all other cases), one key attribute was **response propensity > 0.90**
- Portion (19 percent) of subgroup C was reallocated to subgroup B, based on new definition
- Redefining subgroup categorization could have impact on use of calibration sample outcomes to estimate behavior of main sample

Simulation

Simulation background

Goals

- Formally incorporate **prior knowledge** derived from sample into models to **estimate important survey design parameters**
  - **Sample member response propensities**
  - **Contact propensities**
  - Participation costs
  - Contact costs
- **Update knowledge** as data collection progresses
- Goals are naturally well-suited to Bayesian framework
- Exploratory case study using HSLS:09 data

Models
In this exploratory analysis we focus only on the **response propensity** and **contact models**

Simulation background

BADEN Gibbs sampler

Approach

- Leveraged Gibbs sampler and approach developed by Schouten et al. (2017) to estimate survey design parameters
- Gibbs sampler allows for estimation of joint posterior distributions of interest that cannot be expressed in closed form

- Generate **simulation** data set using
  - Known distribution of auxiliary variables (of interest)
  - Auxiliary variables were kept simple: only **sex** (Female, Male) and **subgroup** (A, B, C)
  - Aggregated sample paradata through phases (e.g., probability of contact, probability of participation)
- **Elicit priors** (informed and uninformed)
- Construct generalized linear models to estimate design parameters of interest
  - Response propensity (probit)
  - Contact propensity (probit)
- **Estimate posterior** regression parameters using Gibbs sampler

Simulation background

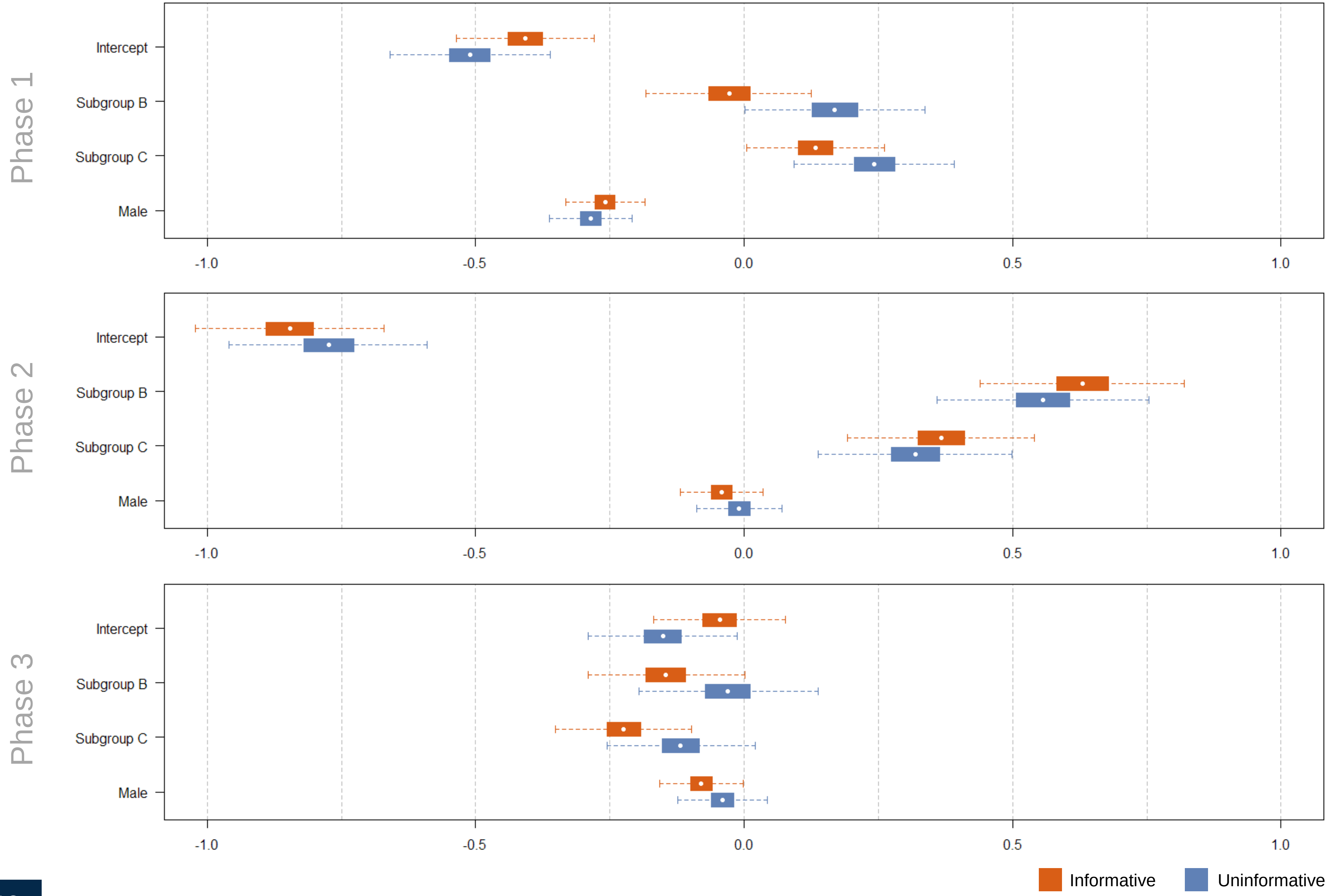
BADEN Gibbs sampler

Priors

Prior distribution for the regression parameters is multivariate normal

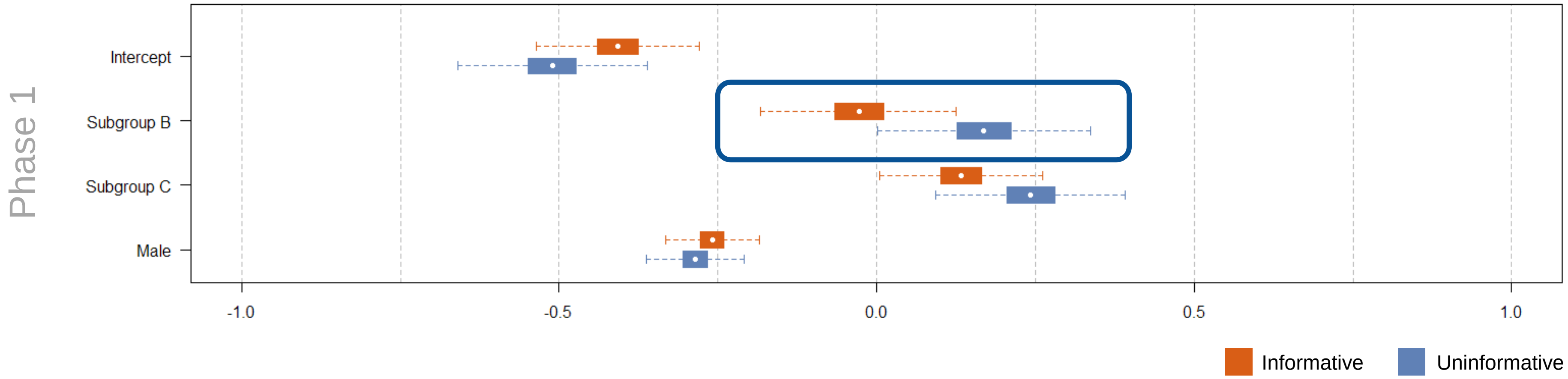
- **Informative prior:** elicited from calibration sample data
- **Uninformative prior:** equivalent variance for all covariates; priors express lack of knowledge at the start of data collection

Posterior response model coefficients



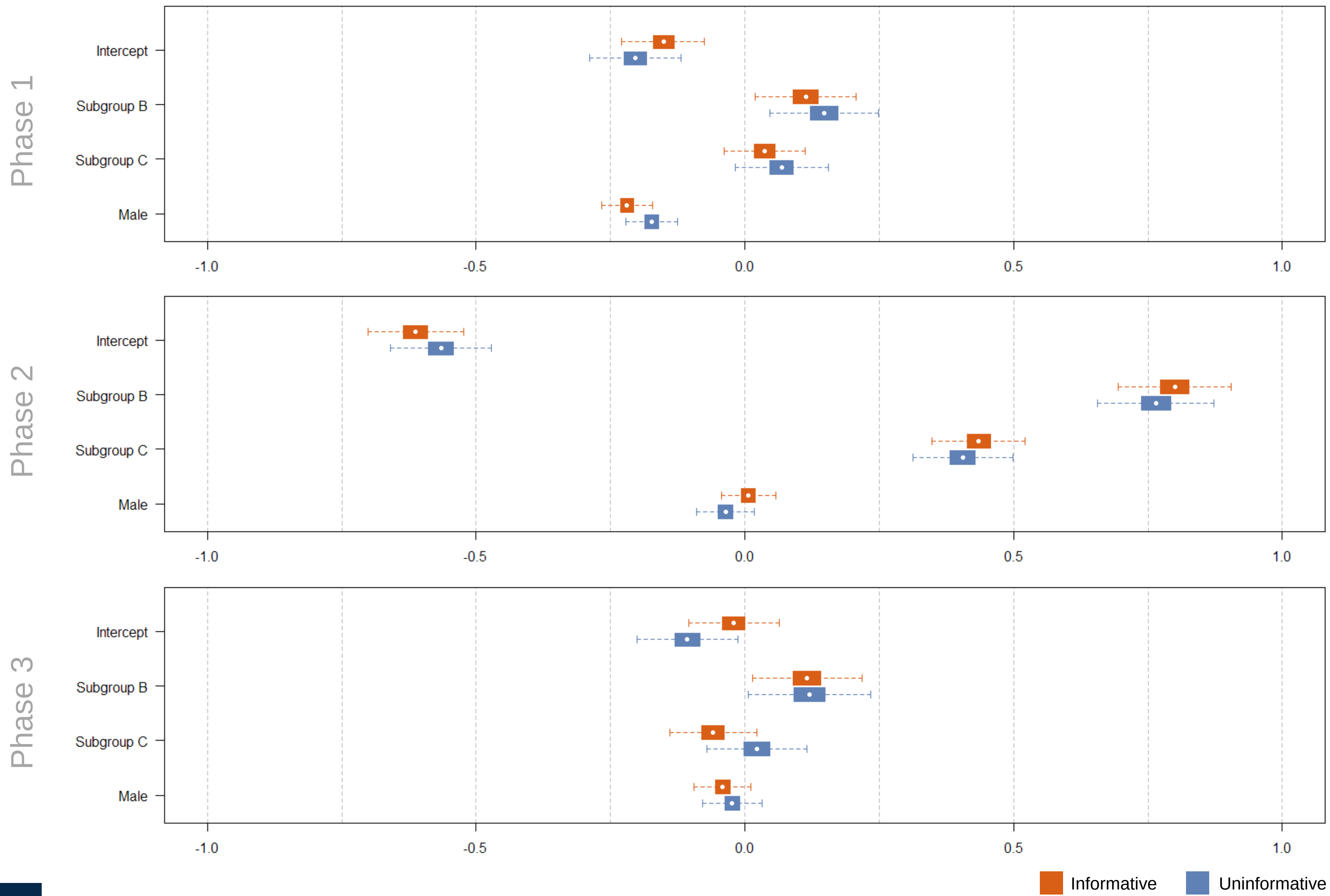
Posterior response model coefficients

Impact of redefining subgroup B



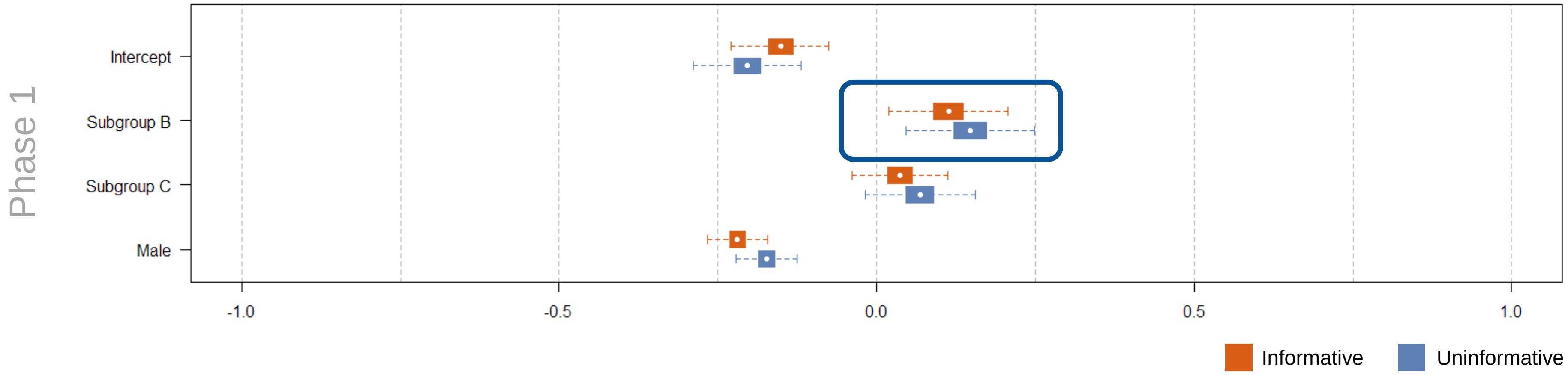
- **Very large differences** on posterior distribution of regression coefficients associated with subgroup (particularly subgroup B), depending on usage of **informative or uninformative priors**
- Suggests calibration-informed prior is misspecified
- Especially marked effect during phase 1, before incentive boosts deployed in phase 2
- Despite efforts to redefine subgroup B with similarly cooperative respondents, changes to prior mid-data collection can have substantive impacts on posterior distributions
- Testing interventions with calibration sample may have caused further misspecification

Posterior contact model coefficients



Posterior contact model coefficients

Impact of redefining subgroup B



- Similar impact for contact parameters not observed
- Subgroup B and subgroup C do not appear to differ in their parameters for contact, but their behavior for response does seem to significantly differ under the calibration study (i.e., informed priors)

Discussion

Calibration-informed priors

- While being responsive to client and resource constrains, adaptations mid-collection can reduce utility of prior with cascading effect on posterior
- Prior and observed data should share the same underlying probability mechanism in order to provide added value
- Making decisions based on the calibration sample with misspecified prior may lead to different actions being applied during the main sample; takes time for effect of prior to be reduced
- Use of calibration study as prior (with goal of experimentally testing incentives, definitions, etc.) vs. use of field test (closer to a survey "dry run") as prior

Next steps

- Testing various calibration sample sizes with various main sample sizes
- Considering methods to formally attach less weight to prior that we know differs. Any utility?
- Using estimates from phase one as priors for phase two, etc. Given phase differences, does this make sense?

Thank you

Contact

Daniel Pratt

djp@rti.org

Michael Duprey

mduprey@rti.org

RTI INTERNATIONAL