Modeling Nonresponse Bias Likelihood and Response Propensity

Daniel Pratt, Andy Peytchev, Michael Duprey, Jeffrey Rosen, Jamie Wescott

Background

- Substantial uncertainty in survey outcomes
- With respect to nonresponse:
  - Current response rates provide potential for nonresponse bias in survey estimates
  - Pursuing the full sample with increased effort is inefficient and often infeasible

Approach

- Identify the main objective
  - Minimize nonresponse bias
- Devise multiple phases of data collection, each altering the data collection protocol
  - Phases should have complementary features (Groves and Heeringa, 2006)
  - Identify which nonresponding cases will likely lead to reduction in nonresponse bias, if interviewed
- Implement the protocols that should increase participation among the identified nonrespondents
- Evaluate results

Identification of Targeted Sample Cases

- Estimate response propensities to identify those most likely to have been excluded from the respondent pool
- Common approach to propensity estimation:
  - Assume everyone has an underlying propensity to respond
  - Use ***all available information*** to estimate the propensity to respond

Key Assumption

- Assumes that the estimated propensities are highly correlated with the survey variables, necessary for the approach to reduce nonresponse bias
- Paradata such as prior round nonresponse and needed level of effort tend to be:
  - Strongly correlated with nonresponse (e.g., Wagner et al., 2014)
  - Weakly correlated with survey measures (e.g., Wagner et al., 2014)
- Could explain why targeting has been ineffective (e.g., Peytchev, Riley, Rosen, Murphy, and Lindblad, 2012)

Proposed Approach

- Devise propensity models that:
  - Deliberately exclude strong predictors of nonresponse but are very weakly associated with survey variables of interest
  - Deliberately identify and select predictors that are highly correlated with the survey variables
- Main objective is not to identify the model that best identifies the response propensities, but to identify which nonrespondendents are likely contributing to nonresponse bias
  - The strong predictors of response propensity could “overwhelm” the correlates of the survey variables in the model
- Let’s name this model a ***bias likelihood model***

High School Longitudinal Study of 2009 (HSLS:09)

- 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)
- Estimate two sets of response propensities:
  - Response propensity model (maximize prediction of second follow-up nonresponse)
  - Bias likelihood model (exclude paradata that are strongly predictive of nonresponse)
- Re-estimate the propensities during data collection

Propensity Models

Response Propensity Model

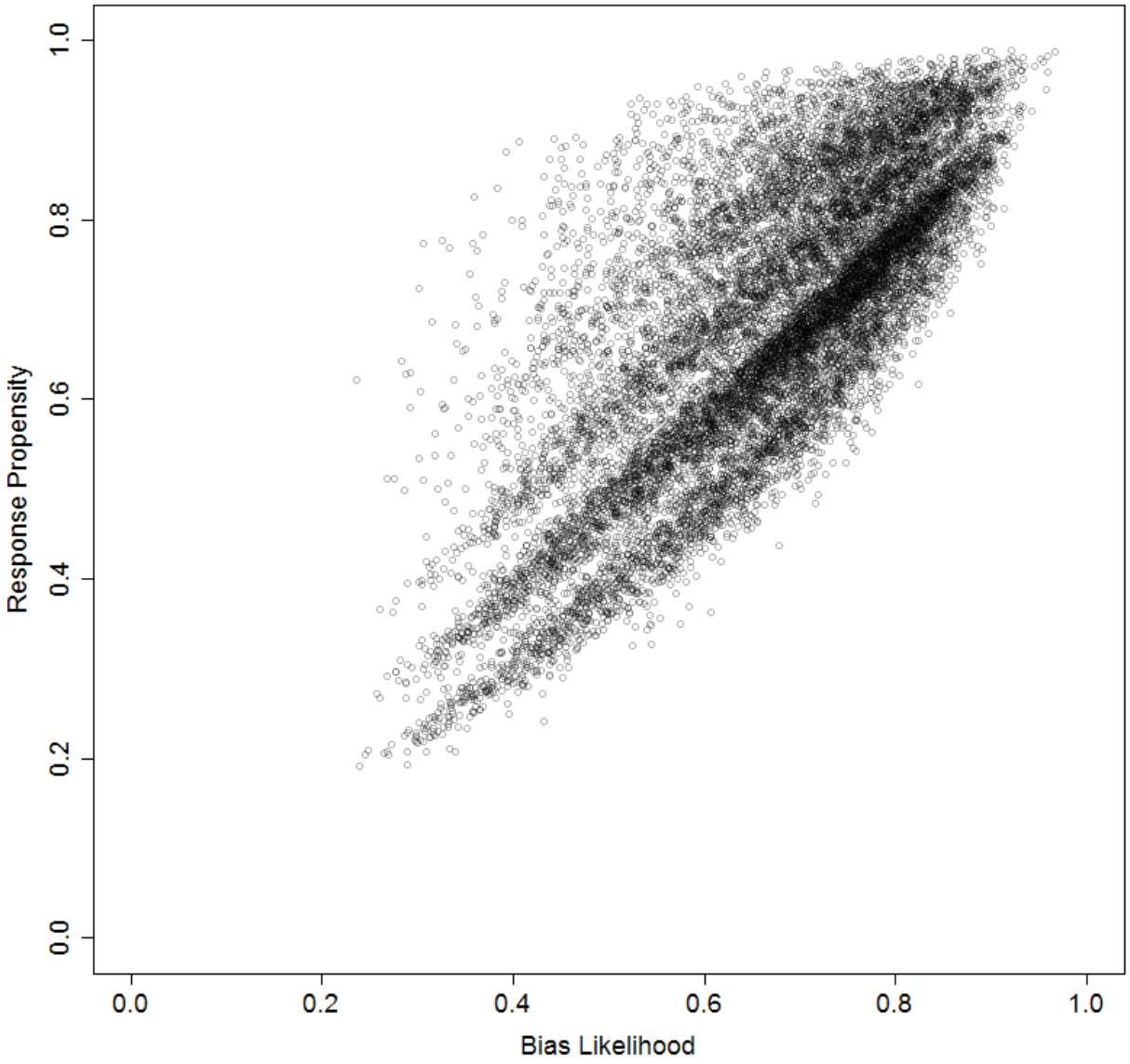
- Estimates unit-level response probability
- **Covariates**
  - Model covariates combine *key variables of interest* (from bias likelihood model) and ***paradata***
- **Dependent variable**
  - Current-round response
- Re-estimated prior to each data collection intervention

Bias Likelihood Model

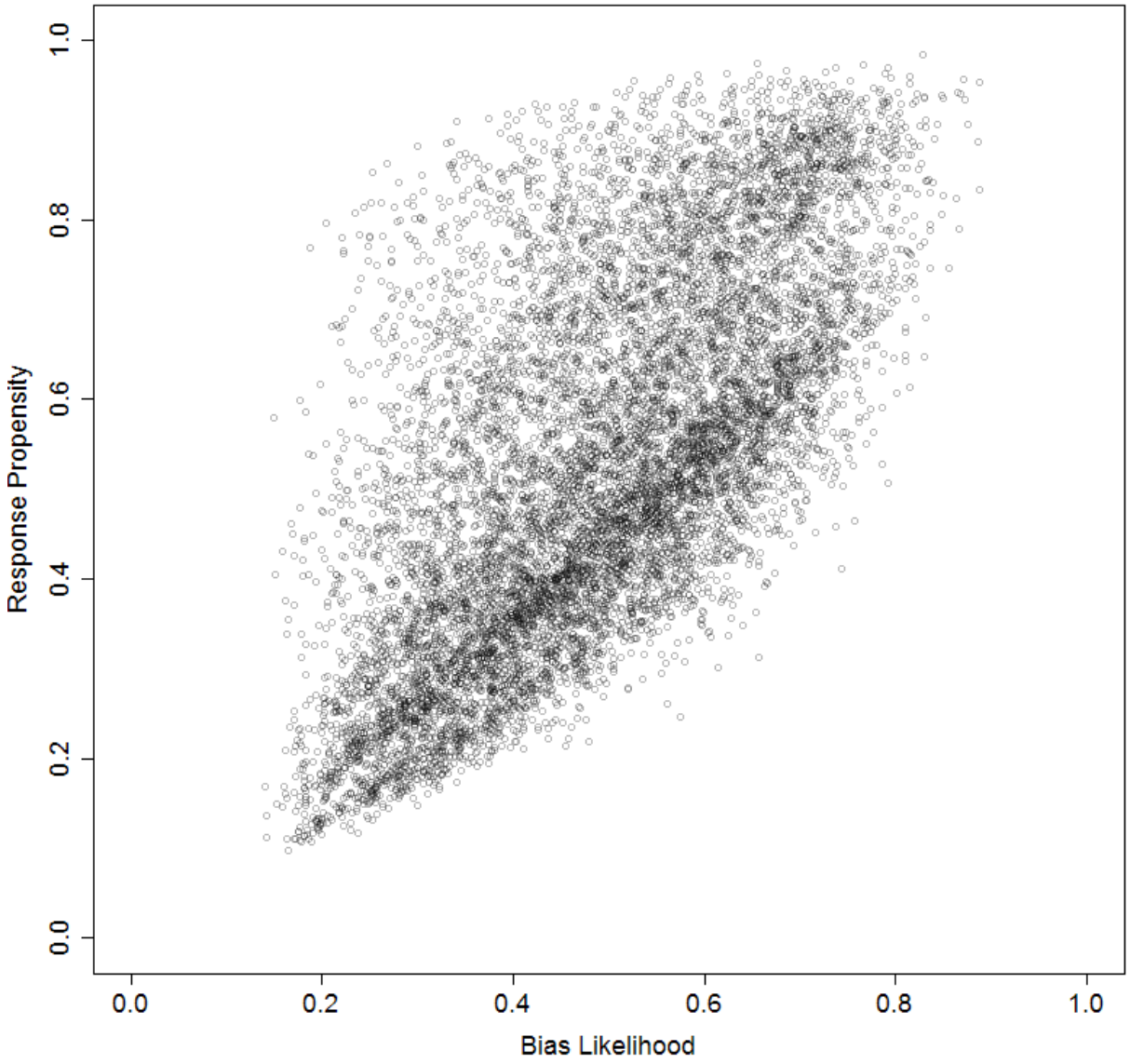
- Identifies nonrespondents in the most underrepresented groups
- **Covariates**
  - Chosen such that differences should proxy nonresponse bias
  - Model ***excludes paradata***
- **Dependent variable**
  - Current-round response
- Re-estimated prior to each data collection intervention

Does including paradata overwhelm bias likelihood model?

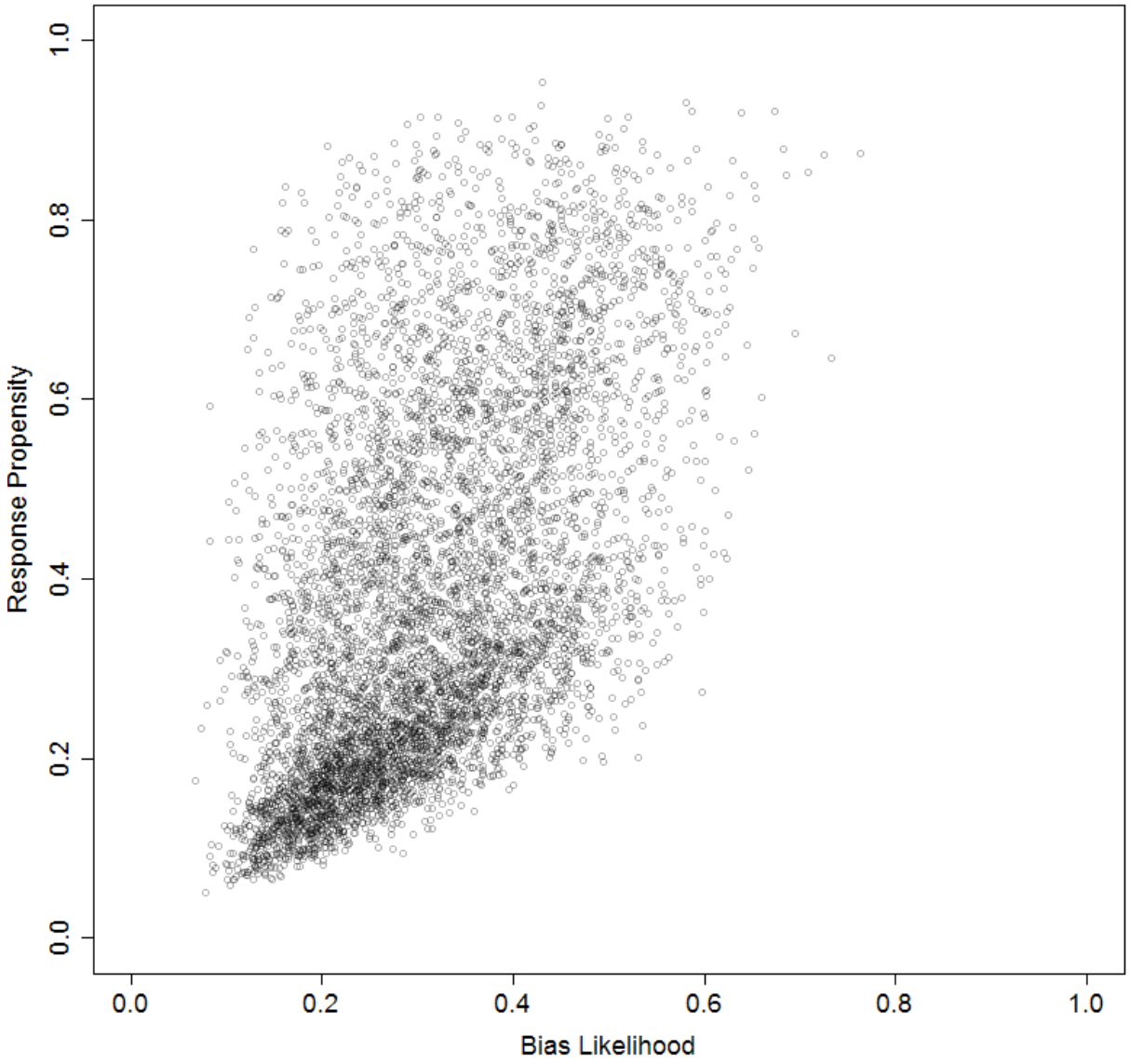
Response Propensity / Bias Likelihood – Start Interventions



Response Propensity / Bias Likelihood – Middle (12 weeks)

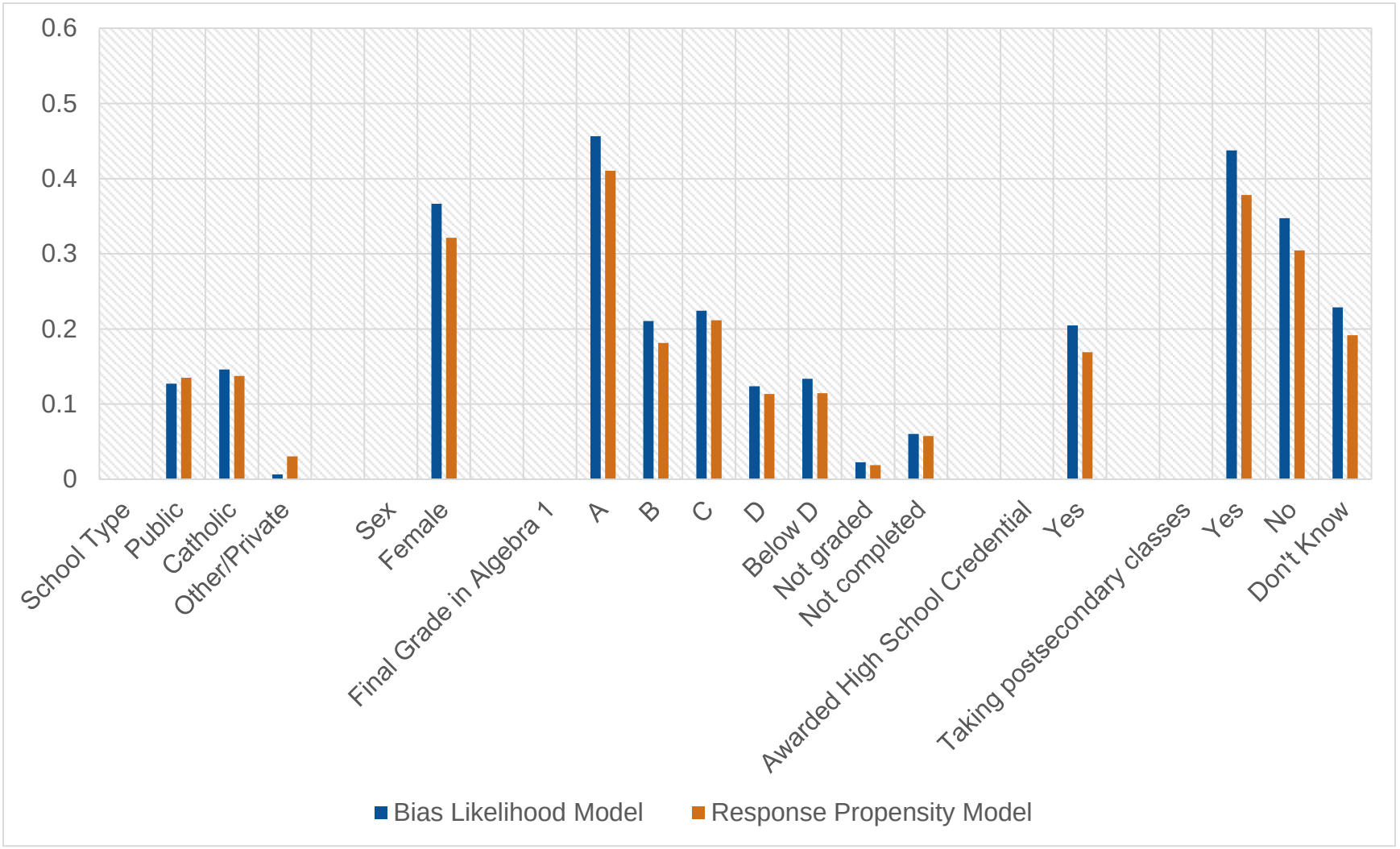


Response Propensity / Bias Likelihood – End (32 weeks)

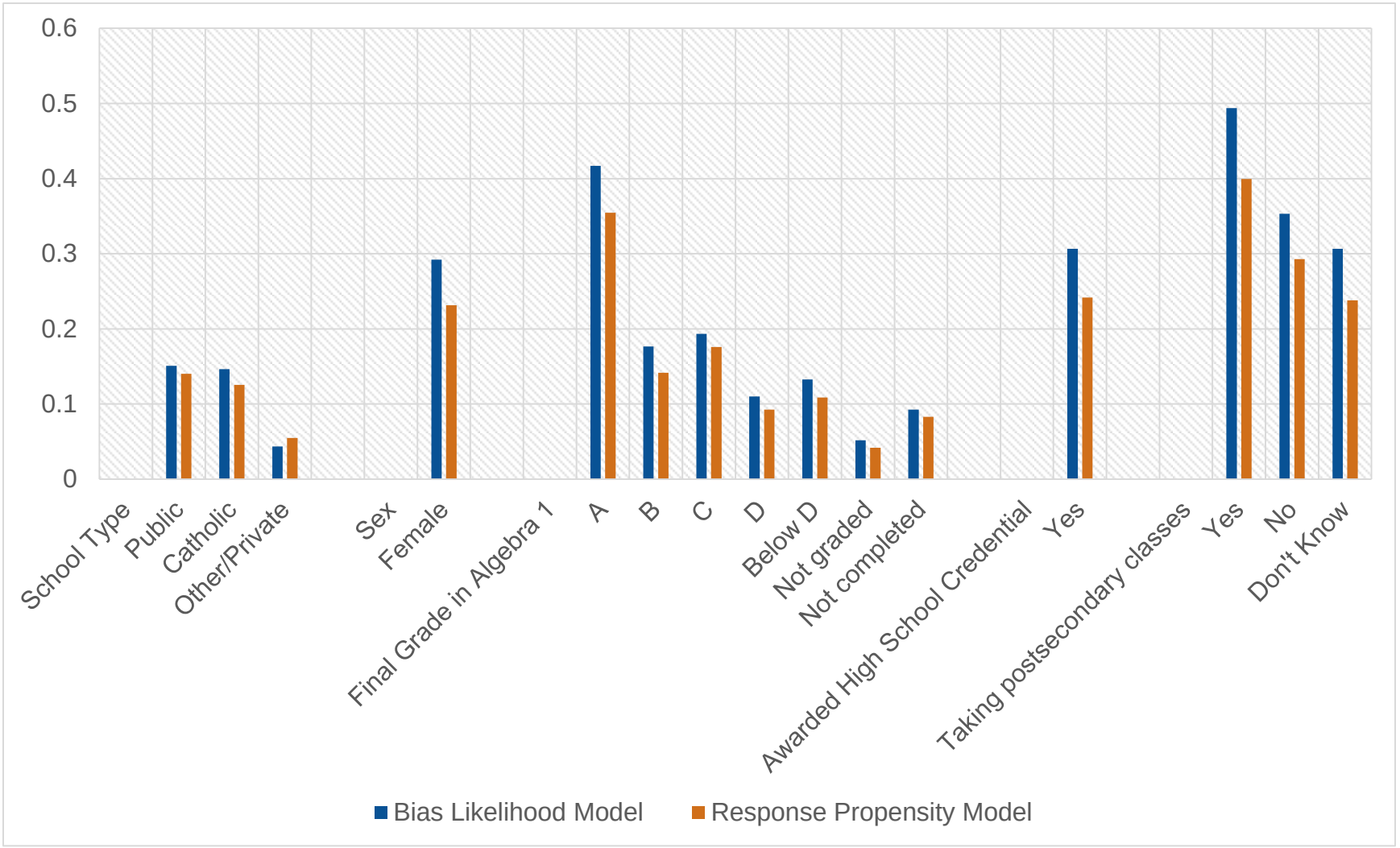


How do the models differ in the estimation of propensities that are associated with survey variables?

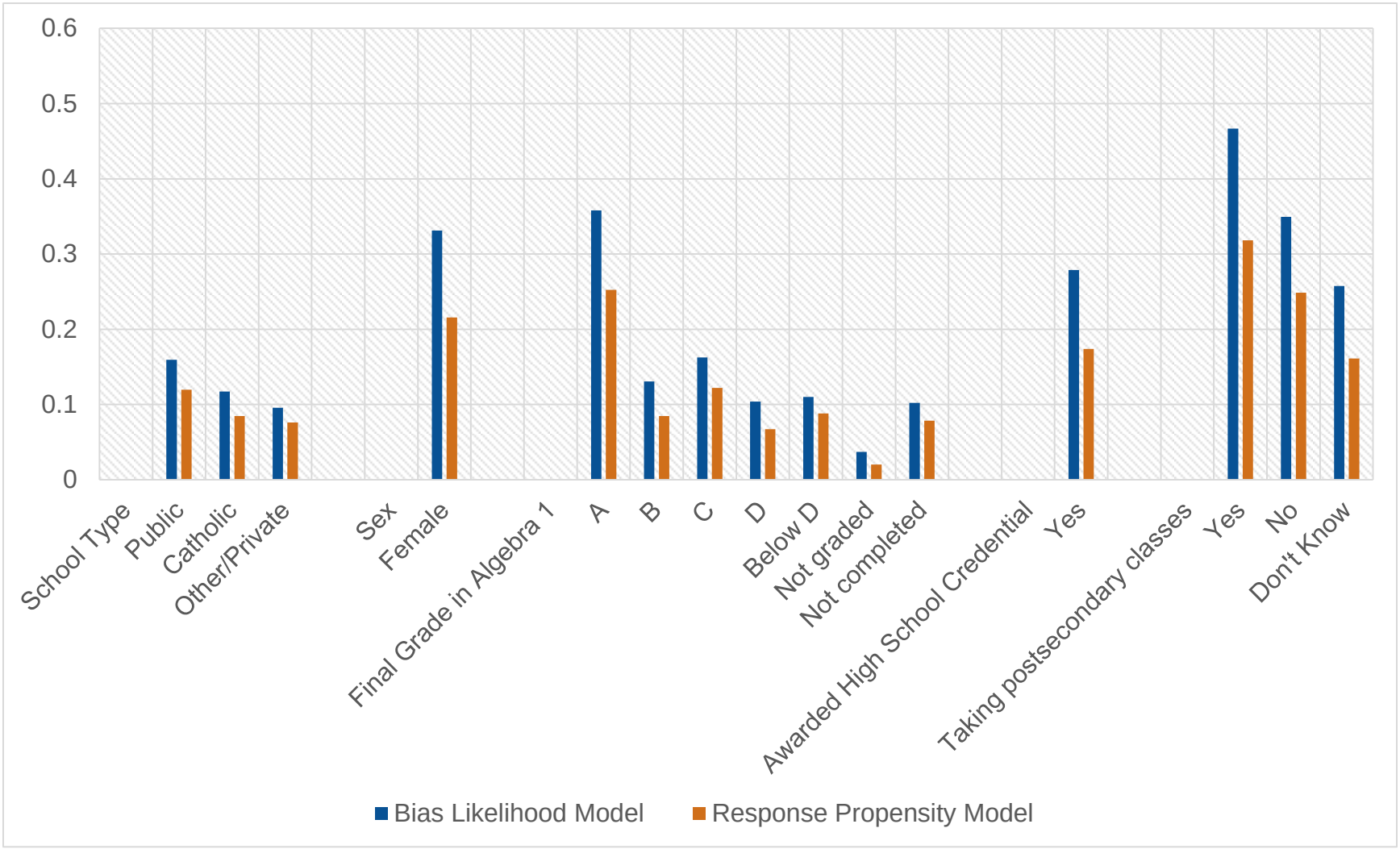
Correlations – Start Interventions



Correlations – Middle (12 weeks)



Correlations – End (32 weeks)



Summary and Conclusions

- Even when the propensity model includes the relevant variables that are associated with the variables of interest, the inclusion of paradata to maximize prediction:
  - Led to higher dispersion of response propensities
  - This produced differences between the predicted propensities of the response propensity model which included paradata and the bias likelihood model that excluded the paradata
  - Reduced the associations between the estimated propensities and the key survey variables
- We recommend going forward with the “Bias Likelihood” model approach for Responsive and Adaptive Design interventions, when using a single model

Next Steps

Develop Bayesian approach

- Advantages (and possible disadvantages) of Bayesian updating of response propensity throughout data collection
- Evaluate impact of informative priors on bias likelihood model
- Integrate cost estimation

Thank You

Daniel Pratt
Education and Workforce Development
RTI International
djp@rti.org