Factors Impacting the Accuracy of Interviewer Observations in the U.S. National Survey of Family Growth (NSFG)*

Brady T. West, Ph.D.
Michigan Program in Survey Methodology
Moving Responsive Design Forward
November 4, 2011

* The 2006-2010 NSFG was carried out under a contract with the CDC's National Center for Health Statistics, Contract # 200-2000-07001.

Why Interviewer Observations?

- Hard to find auxiliary variables for post-survey nonresponse adjustments associated with both *Y* (key variables) and *P* (response propensity)
- Interviewers are the eyes and ears of the survey organization in the field, and can be asked to observe selected characteristics related to both *Y* and *P* for the full sample
- Unfortunately, observations are typically judgments and estimates, making them prone to error
- **TSE Framework:** Does reduced quality of the observations lead to estimates with reduced quality?

Key Gaps in the Existing Literature

- Few existing studies have directly examined the error properties of interviewer observations, largely due to a lack of validation data
- No studies to date of implications of error in the interviewer observations for post-survey nonresponse adjustment of estimates
- **No studies to date of factors impacting the accuracy of interviewer observations in a face-to-face survey**
- **No studies to date of effective observational strategies used by face-to-face interviewers**
- Existing methods for nonresponse adjustment fail to account for possible errors in auxiliary variables

Quality of Two NSFG Observations (West, 2011, submitted)

- When using actual survey responses for validation, interviewers were 78% accurate when judging a *behavioral* trait (current sexual activity)
- FPR of 0.566, FNR of 0.119 for sexual activity judgments (systematic false positives!)
- When using household roster information for validation, interviewers were 72% accurate when judging a *household* trait (presence of kids)
- FPR of 0.169, FNR of 0.557 for observations on presence of children (systematic false negatives!)

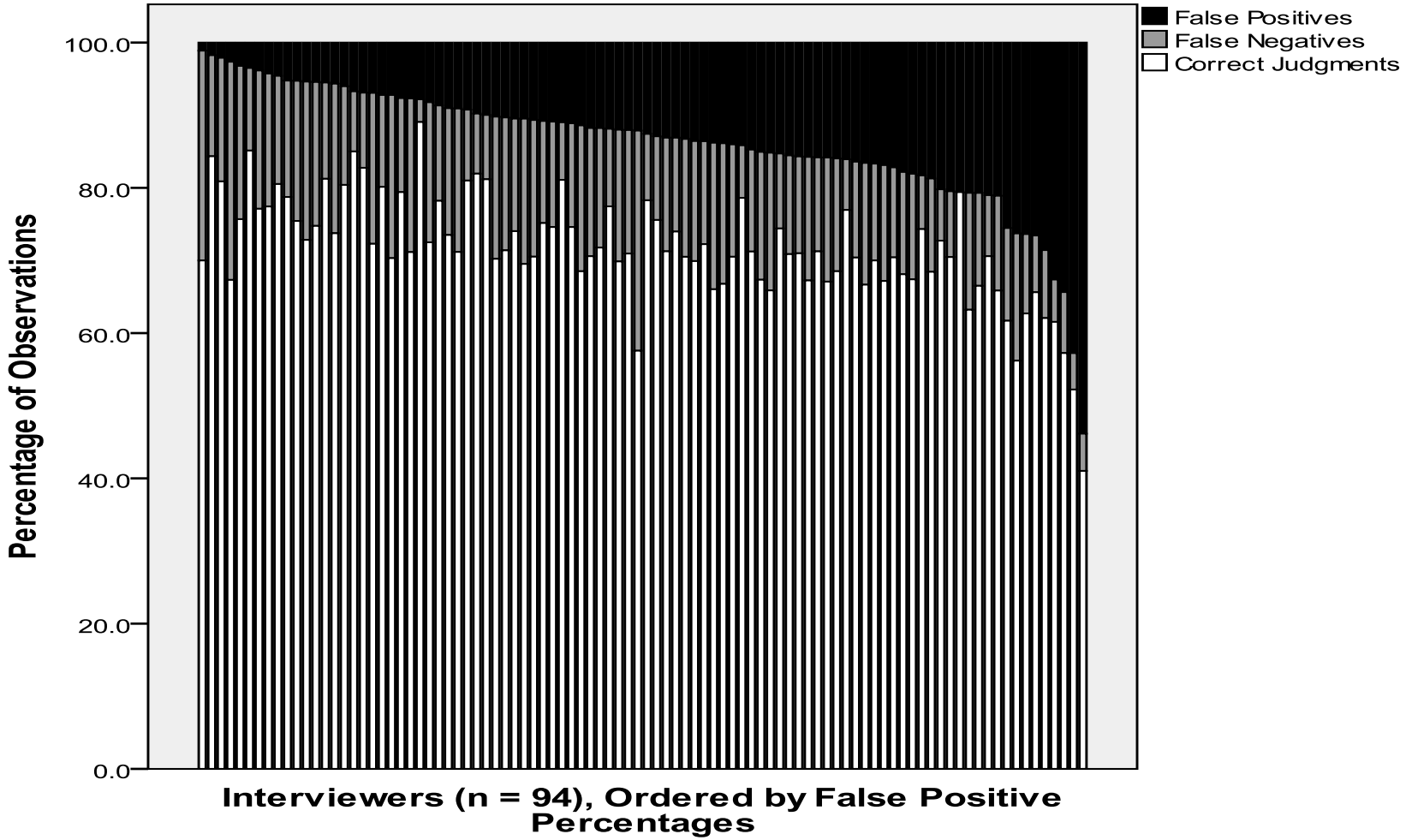
Implications of Error (West, 2011, submitted)

- Small simulation study based on artificial population of all female respondents in NSFG Cycle 7
- **Population variables:** true response on sexual activity, interviewer judgment on sexual activity, parity, and number of partners in past year
- 1,000 simulated samples of n = 500, with response on parity and partners simulated as a function of "true" sexual activity

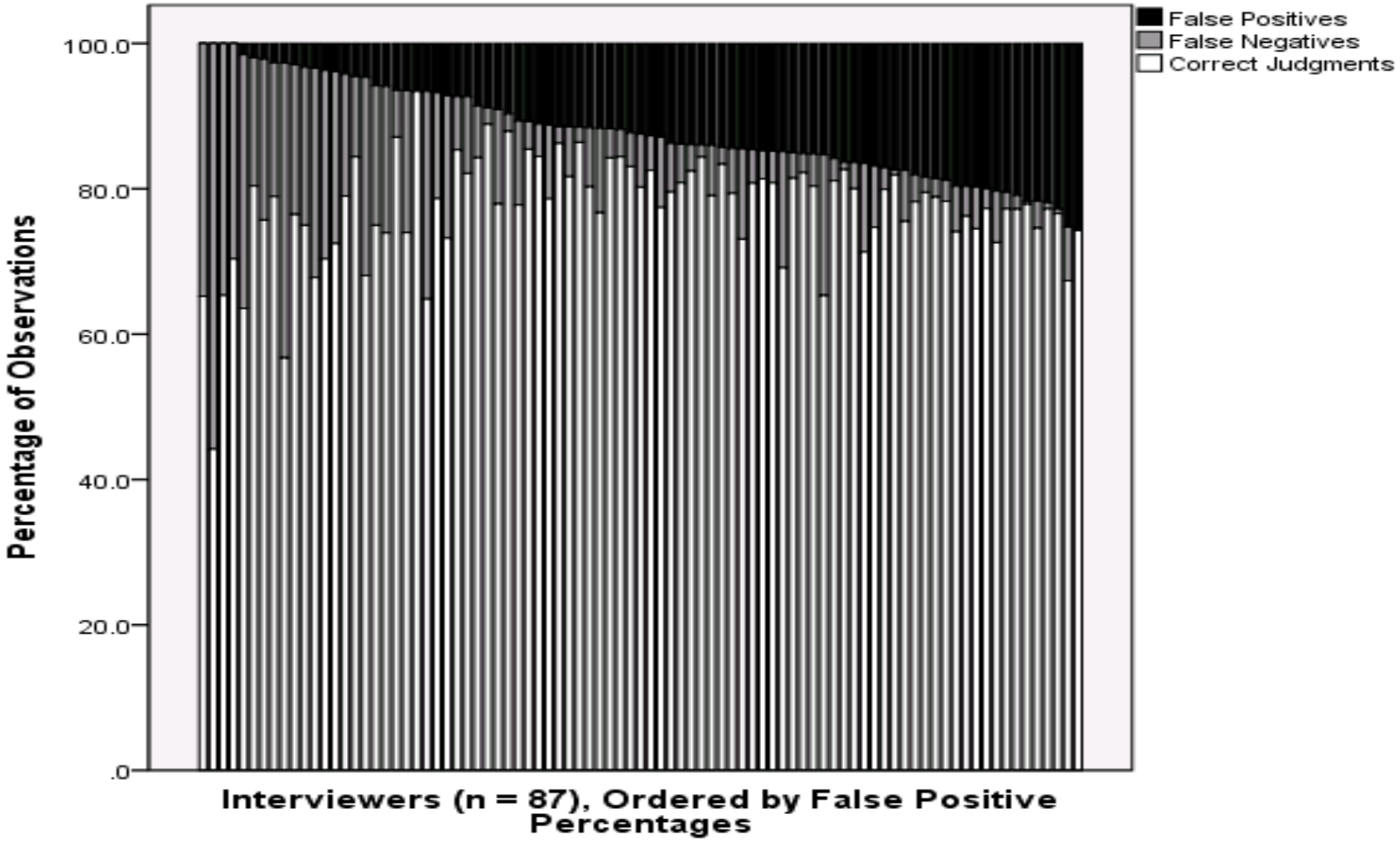
Implications of Error, cont'd

- **Adjusted estimates for mean # partners (strong association with "true" sexual activity) based on judgments had *higher* relative bias than CC estimates, and similar (low) coverage**
- **Why? (Lessler and Kalsbeek, 1992)**
  - R had a higher mean # partners than NR in both classes
  - The class defined by a judgment of not currently sexually active had a *higher* mean # partners
  - The response rate was higher in the class judged to be sexually active
- Results may not be as severe for subgroups
- Also Biemer et al. (2011): differential error in interviewer-reported counts of calls depending on disposition can substantially bias estimates

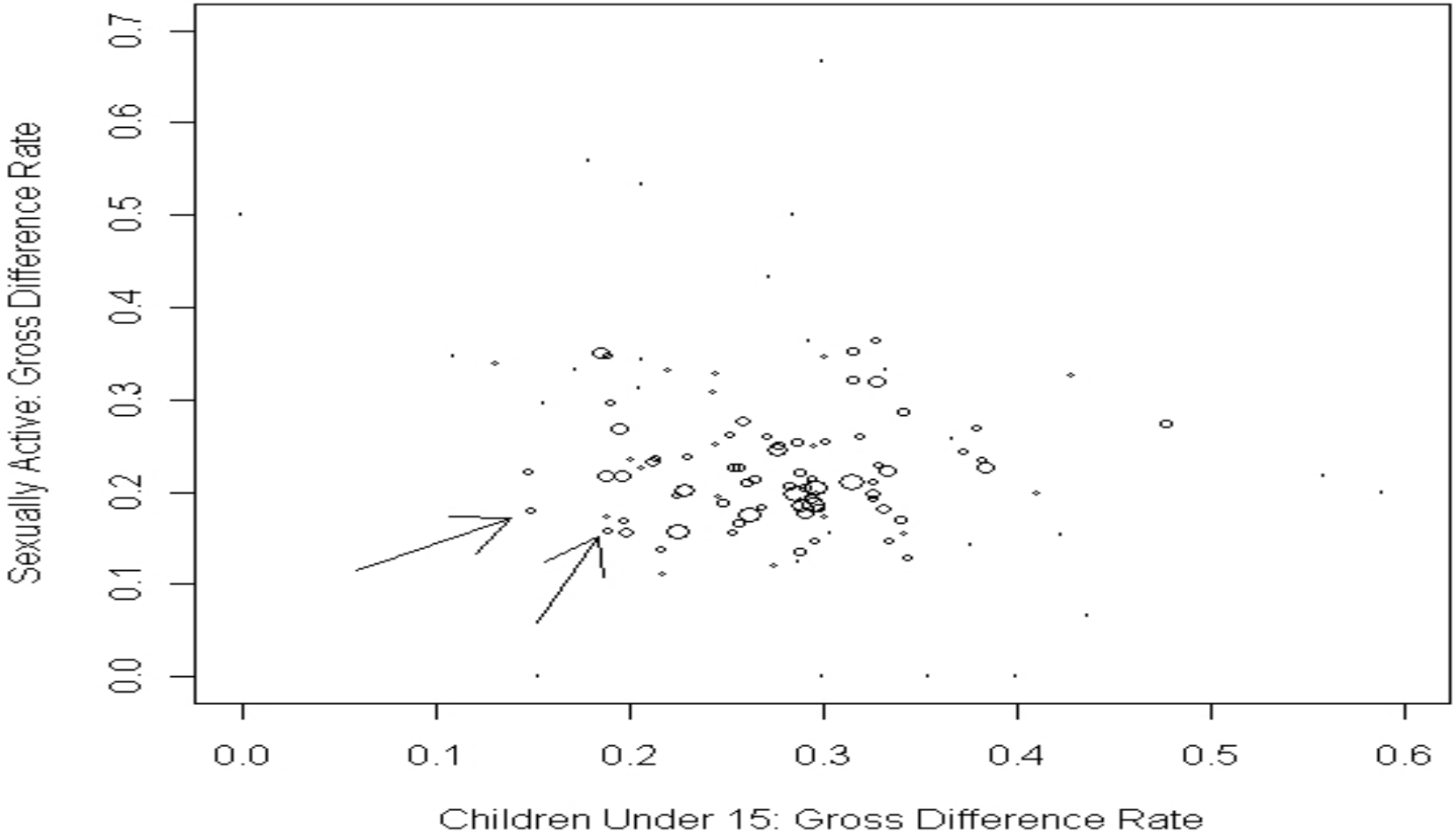
Interviewer Variance in Accuracy of Judgments on Presence of Children



Interviewer Variance in Accuracy of Judgments on Current Sexual Activity



Are Accuracy Rates Correlated?



Research Questions

1. In multilevel multinomial logistic regression models, what are respondent- and interviewer-level factors that impact that accuracy of the two interviewer observations in the NSFG?
2. Does a theory-driven design strategy for improving observation accuracy actually increase accuracy when controlling for other factors at both levels?
3. Do interviewers vary in terms of observational strategies used in the field?
4. Do varying observational strategies lead to varying levels of accuracy in the observations?

Theoretical Expectations from the Social Psychology Literature

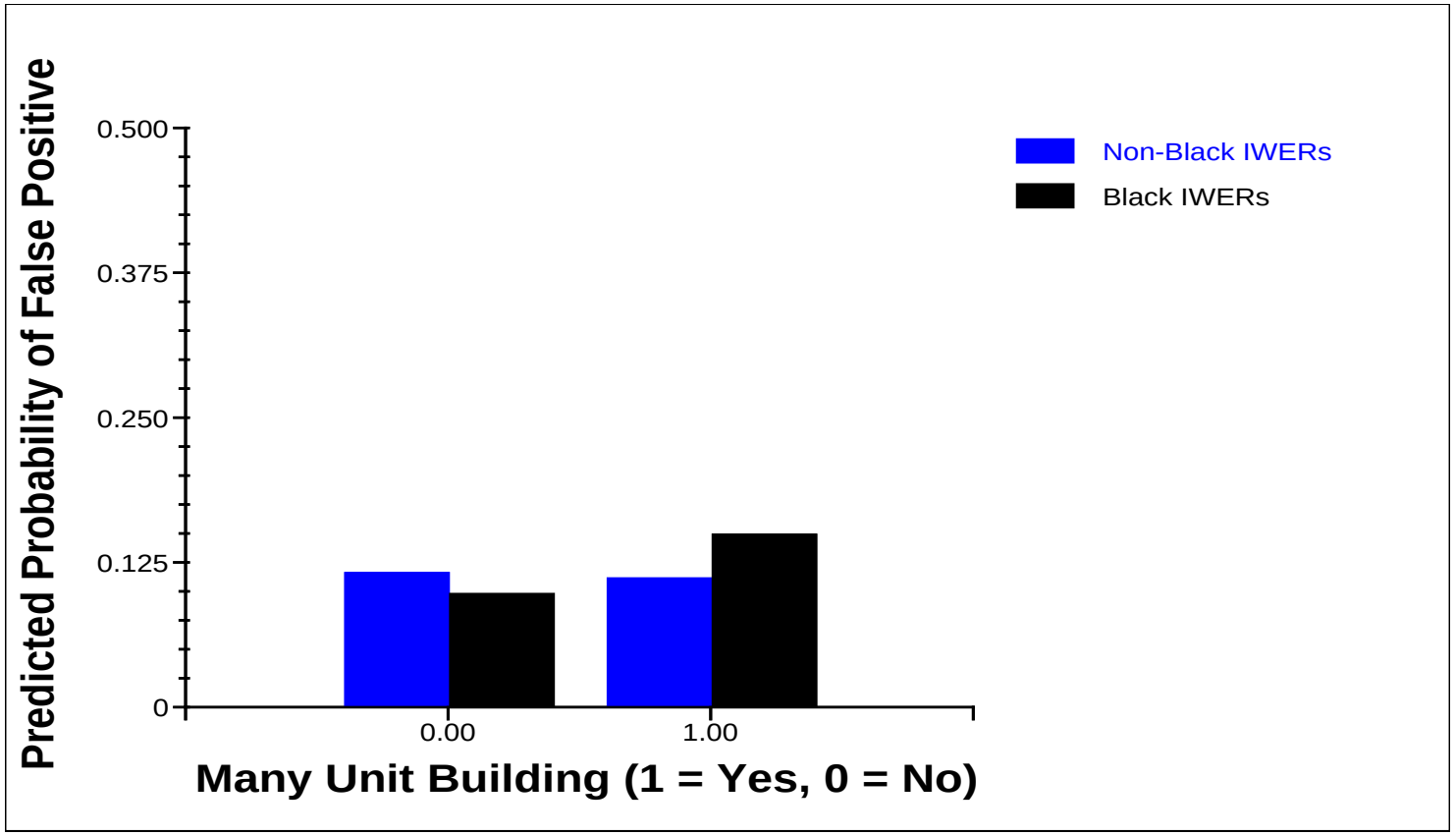
- The difficulty of the observational task rather than individual ability will influence accuracy
- Interviewers with features relevant to the judgments will have improved accuracy
- Providing interviewers with a set of features predictive of the features being observed and available to their observation will help to improve observation accuracy

Results

- **Question 1:**
  - Accuracy on the household judgment (presence of kids) was a function of respondent-level features (e.g., urban areas had reduced accuracy) and interactions between respondent- and interviewer-level factors
  - Accuracy on the behavioral judgment (sexual activity) was a function of independent effects of respondent- and interviewer-level factors; no significant interactions
  - **Significant unexplained variance remained among interviewers in false-positive and false-negative logits**; also large *negative covariances* between random effects in the two logits, indicating the systematic nature of the errors

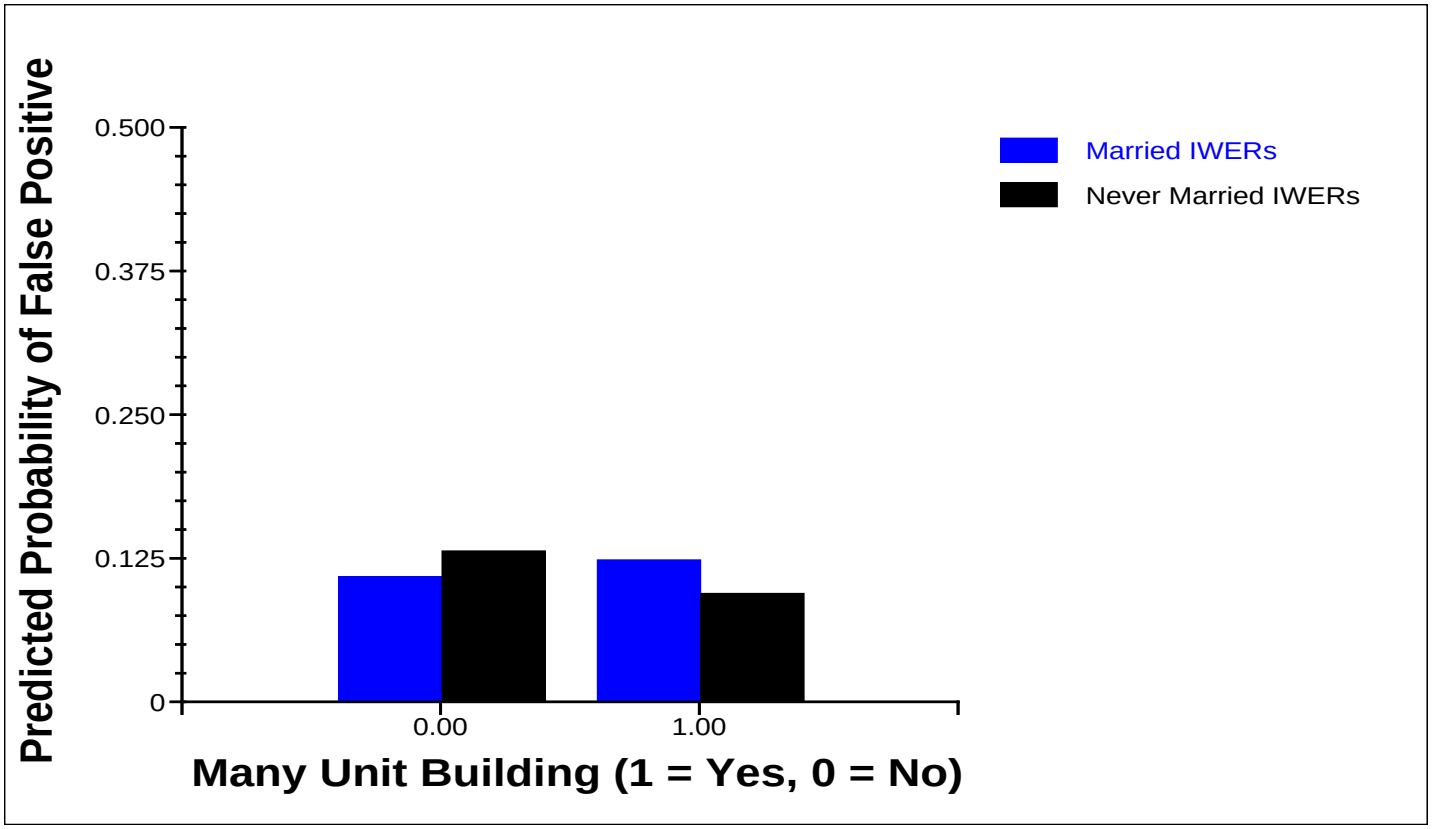
Results: Presence of Children

The relationship of many-unit buildings with the probability of a false positive varies as a function of interviewer ethnicity…



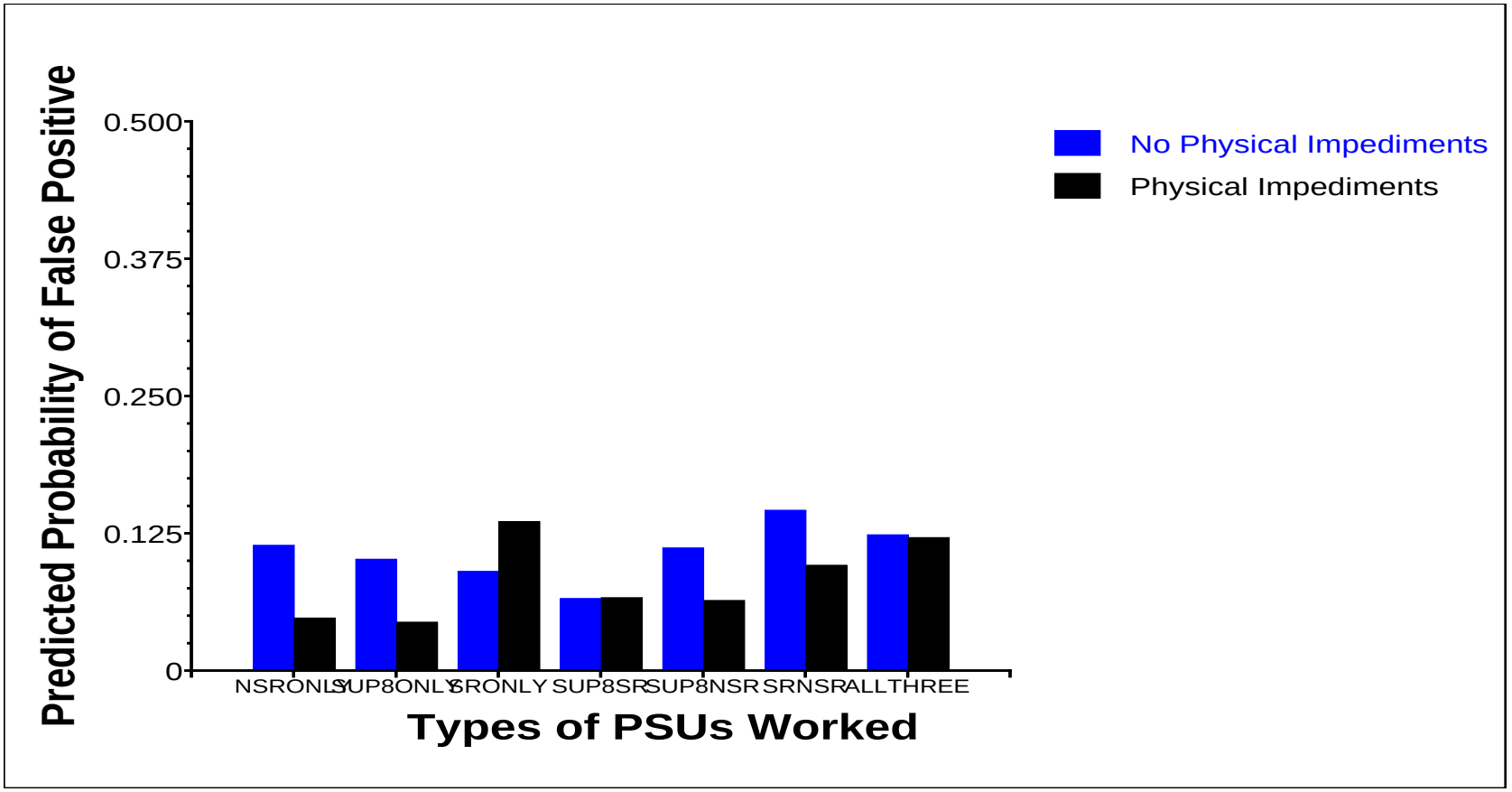
Results: Presence of Children

Marital status also moderates the impact of many-unit buildings…



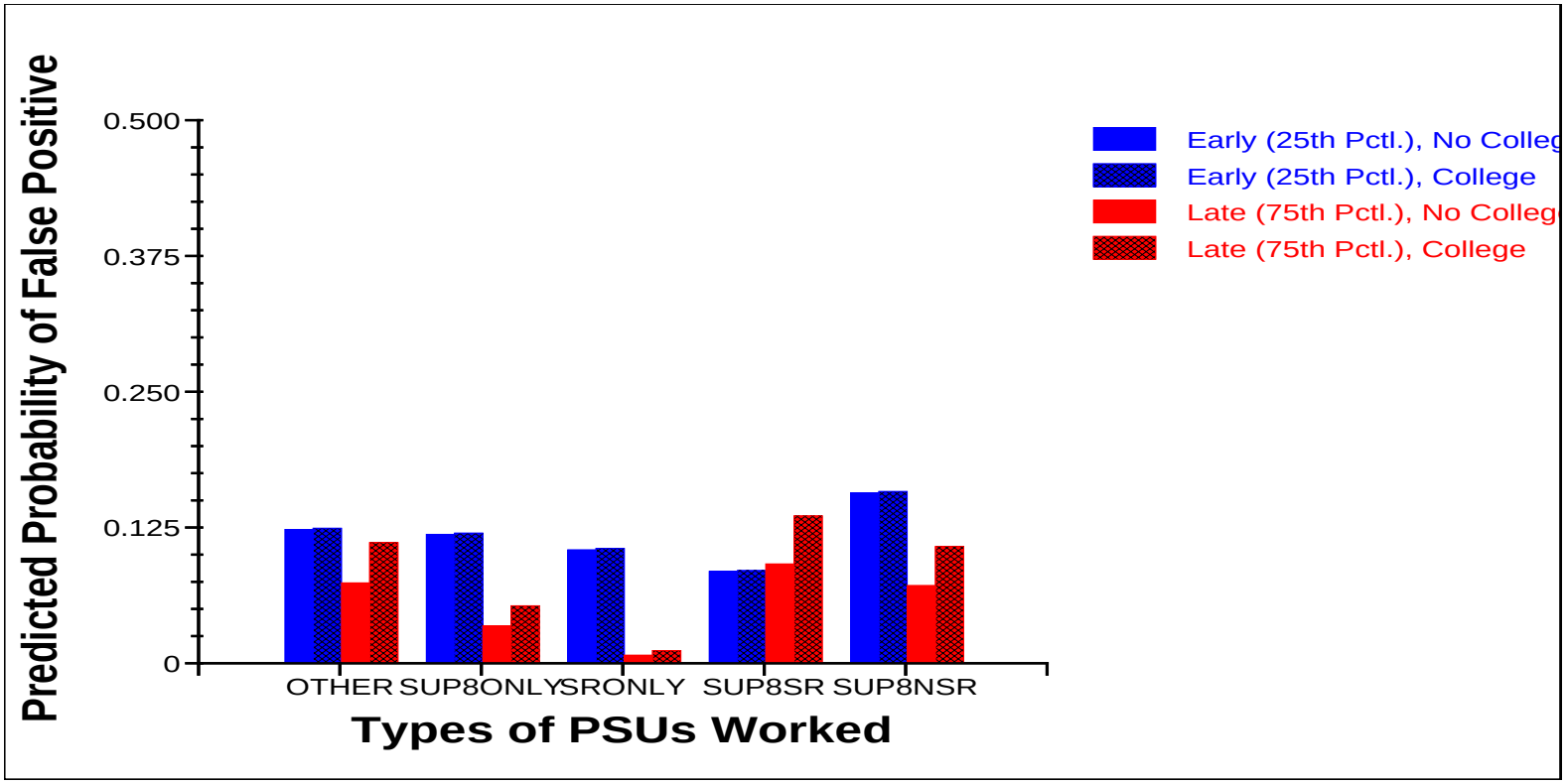
Results: Presence of Children

The impact of physical impediments to access varies depending on the types of PSUs worked by the interviewers…



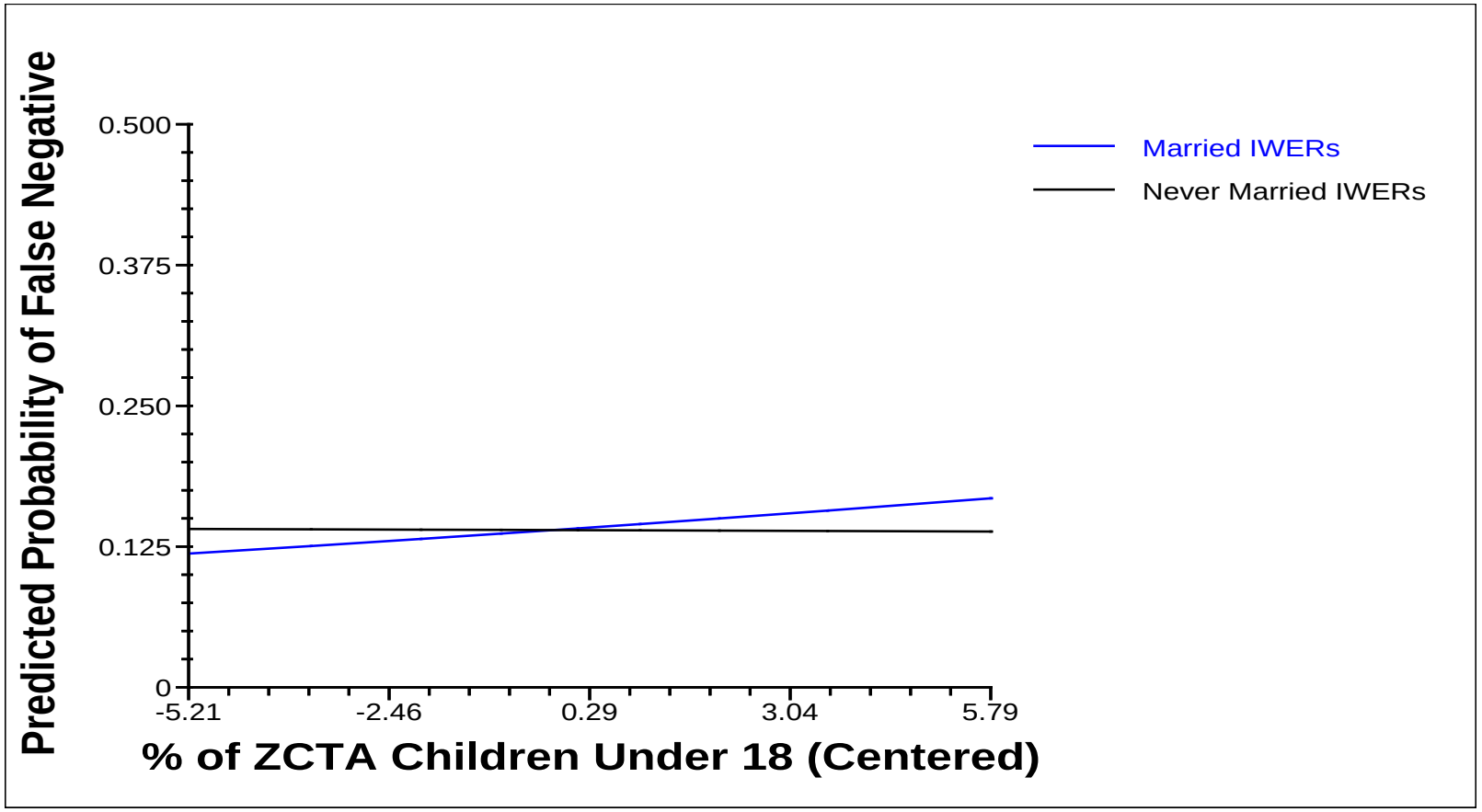
Results: Presence of Children

A three-way interaction: experience helps except for those interviewers working in both Super 8 and SR PSUs; college education *hurts* later on in the data collection, but not earlier on



Results: Presence of Children

The relationship of the % of children in the zip code tabulation area with the probability of a false negative varies depending on marital status



Results, cont'd

- **Question 2:**
  - A respondent-level indicator of measurement in Quarters 15 and 16 (when interviewers were first provided with significant predictors of sexual activity) was found to significantly reduce the odds of a false positive relative to a correct judgment
  - This result held when controlling for amount of Cycle 7 experience (in # days since starting) and all other respondent- and interviewer-level factors
  - Evidence in support of this design strategy, given documented FP problems with this observation

Results, cont'd

- **Question 3:**
  - 3,992 interviewer justifications for sexual activity judgments were coded on 13 indicators of reasons mentioned (e.g., age, relationship status), along with the # of words used in the justification
  - Indicators aggregated to interviewer level (13 percentages and one mean), and standardized
  - Cluster analysis of aggregate indicators revealed four distinct clusters of interviewers, varying in terms of strategies used (e.g., focus on age only)

Examples of Justifications

"He works and goes to school and lives here with his twin - I don't think he could have someone over as the carpet is all taken up and it smells badly of dog poo."

"He has a tattoo `Carol` over his heart."

"She did not appear to be very world wise; her appearance was not well kept and she advised that she had been home schooled since 10 grade which should have been the beginning of her experimental time; she stays at home with a baby all day and has no car."

Examples of Justifications, cont'd

"Said no one lives with him right now which indicates someone has been living here; also he was sleeping on sofa bed in LR leading me to believe whoever just moved out took the bed."

"College student away, 19, affluent background, educated parents, had big social life, his house was party central for the neighborhood per mother."

"R selected Mr E was present during screening, Mr E seems happy to have extra cash in hand because said he could invite a lady friend out."

Four Clusters of Interviewers

Cluster	*Features*	*GDR*	*FPR*	*FNR*
1 (n = 20)	Focus on Living Arrangement and Household Features	0.247	0.413	0.196
2 (n = 7)	Focus on Appearance, Personality, and Age	0.191	0.536	0.087
3 (n = 11)	**Focus on Relationship Status and "Hunches"**	**0.168**	**0.515**	**0.070**
4 (n = 5)	Primary Focus on Age and Little Else	0.171	0.795	**0.006**

Results, cont'd

- **Question 4:**
  - Separate multilevel multinomial models of accuracy on the sexual activity judgment were fitted in Quarters 15 and 16, when justifications for the observations were collected
  - Indicators for three of the four clusters did not explain any variance among interviewers in the intercepts in the false positive logit (when controlling for the same other factors)
  - 16% of the unexplained variance in the intercepts in the false negative logit was explained by the indicators
  - The cluster (4) focusing primarily on age had reduced odds of a false negative relative to a correct judgment

Implications for Practice

- Multilevel modeling can identify respondent- and interviewer-level factors that impact the accuracy of interviewer observations (given validation data)
- Providing interviewers with observable predictors of the features with which they are tasked with observing can help to improve observation accuracy
- Specific results can be used to identify particular combinations of factors that result in *difficult* observations (e.g., married interviewers and many-unit buildings when judging presence of children) or *higher accuracy* observations

Implications for Practice, cont'd

- Could replace error-prone observations with model-based predictions or possibly commercially available auxiliary variables
- **Constant communication with interviewers is very important for understanding good strategies!**
- Results from these analyses could be used to understand variance in observational strategies, how that variance could impact accuracy, and how to best standardize observations in future data collections

Future Research Directions

- More research is needed to understand the unexplained variance in accuracy among interviewers (perceptive ability? mood?); many PSU-level features were accounted for in this study
- **Possible intervention study:** does targeting a random subset of interviewers with unusual EBLUPs (based on these models) improve their observation accuracy over time relative to others?
- Randomized interventions are needed to further assess the proposed design strategy (ongoing work)
- Additional qualitative research is needed to understand effective observational strategies in other survey contexts (ongoing discussions with PASS survey interviewers **this week!**)

Overall Conclusions

- Future research directions need to consider the broader implications of errors in auxiliary variables for a variety of survey methodologies aside from nonresponse adjustment (e.g., responsive design)
- If more training is dedicated to improving observations, is there a fair cost-quality tradeoff between requiring the collection of observations and using them for estimation purposes? Are we really achieving gains in the quality of estimates or not?
- Constant monitoring of the quality of observations and factors impacting the quality will only benefit survey agencies using this practice
- Systematic feedback for interviewers may also help!

Questions?

- Thank you for attending!
- Please email me (bwest@umich.edu) with any additional questions, requests for papers or presentations, or citation inquiries!