The Nature of the Bias When Studying Only Linkable Person Records:
Evidence from the American Community Survey

Adela Luque (U.S. Census Bureau)

Brittany Bond (U.S. Department of Commerce)

J. David Brown & Amy O'Hara (U.S. Census Bureau)

FedCASIC March 2014

Disclaimer

- Any opinions and conclusions expressed herein are those of the authors and do not necessarily reflect the views of the U.S. Census Bureau
- All results have been reviewed to ensure that no confidential information on individual persons is disclosed

Overview

- Motivation
- Objectives
- Data & Methodology
- Background on Anonymous Identifier Assignment Process
- Expected Effects
- Results
- Conclusions

Motivation

- Record linkage can enrich data, improve its quality & lead to research not otherwise possible - while reducing respondent burden & operational costs
- Linking data requires common identifiers unique to each record that protect confidentiality
- Census Bureau assigns Protected Identification Keys (PIKs) via a probabilistic matching algorithm: PVS (Personal Identification Validation System)
- Not possible to reliably assign a PIK to every record, which may introduce bias in data analysis

Objectives

- What characteristics are associated with the probability of receiving a PIK? That is, what is the nature of the bias introduced by incomplete PIK assignment?
- Help researchers understand nature of bias, interpret results more accurately, adjust/reweight linked analytical dataset
- Examine bias using regression analysis - before & after changes in PVS. Do alterations to PVS improve PIK assignment rates as well as reduce bias?
  - NORC (2011) described some demographic and socio-economic characteristics of those records not getting a PIK

Data & Methodology

- 2009 & 2010 American Community Survey (ACS) – processed through PVS
  - Ongoing representative survey of the U.S. population
  - Socioeconomic, demographic & housing characteristics
  - 50 states & DC - Annual sample approximately 4.5 million person records
- Probit model for 2009 and 2010 separately
  - Dependent variable = 1 if person record received a PIK (0 otherwise)
  - Covariates:
    - Demographic characteristics: age, sex, race and Hispanic origin
    - Socio-economic characteristics: employment status, income, poverty status, marital status, level of education, public program participation, health insurance status, citizenship status, English proficiency, military status, mobility status, and household type
    - Housing and address-related characteristics: urban vs. rural, type of living quarter, age of living quarter
  - ACS replicate weights
  - Report marginal effects
- 2009 & 2010 results compared – before & after changes to PVS

Background on PVS

- Probabilistic match of data from an incoming file (e.g., survey) to reference file containing data from the Social Security Administration enhanced with address data obtained from federal administrative records
- If a match is found, person record receives a PIK or is “validated”

Background on PVS

- Initial edit to clean & standardize linking fields (name, dob, sex & address)
- Incoming data processed through cascading modules (or matching algorithms)
- Only records failing a given module move on to the next
- Impossible to compare all records in incoming file to all records in reference file → "blocking"
  - Data split into blocks/groups based on exact matches of certain fields or part of fields – probabilistic matching within block

Background on PVS

- 2009 PVS Modules
  - *Verification* – Only for incoming files w/ SSNs
  - *Geosearch* looks for name/dob/gender matches after blocking on an address or address part (within 3-digit ZIP area)
  - *Namesearch* looks for name & dob matches within a block based on parts of name/dob
- Each module has several 'passes' – different blocking & matching strategies

Background on PVS

- 2010 PVS Enhancements
  - *ZIP3 Adjacency Module* looks for name/dob/gender/address matches after blocking on address field parts in areas adjacent to 3-digit ZIP area
  - *DOB Search Module* looks for name/gender/dob matches after blocking on month & day of birth
  - *Household Composition Search Module* looks for name/dob matches for unmatched records that are seen in past at same address with PIKed record
  - *Inclusion of ITINs in reference file*

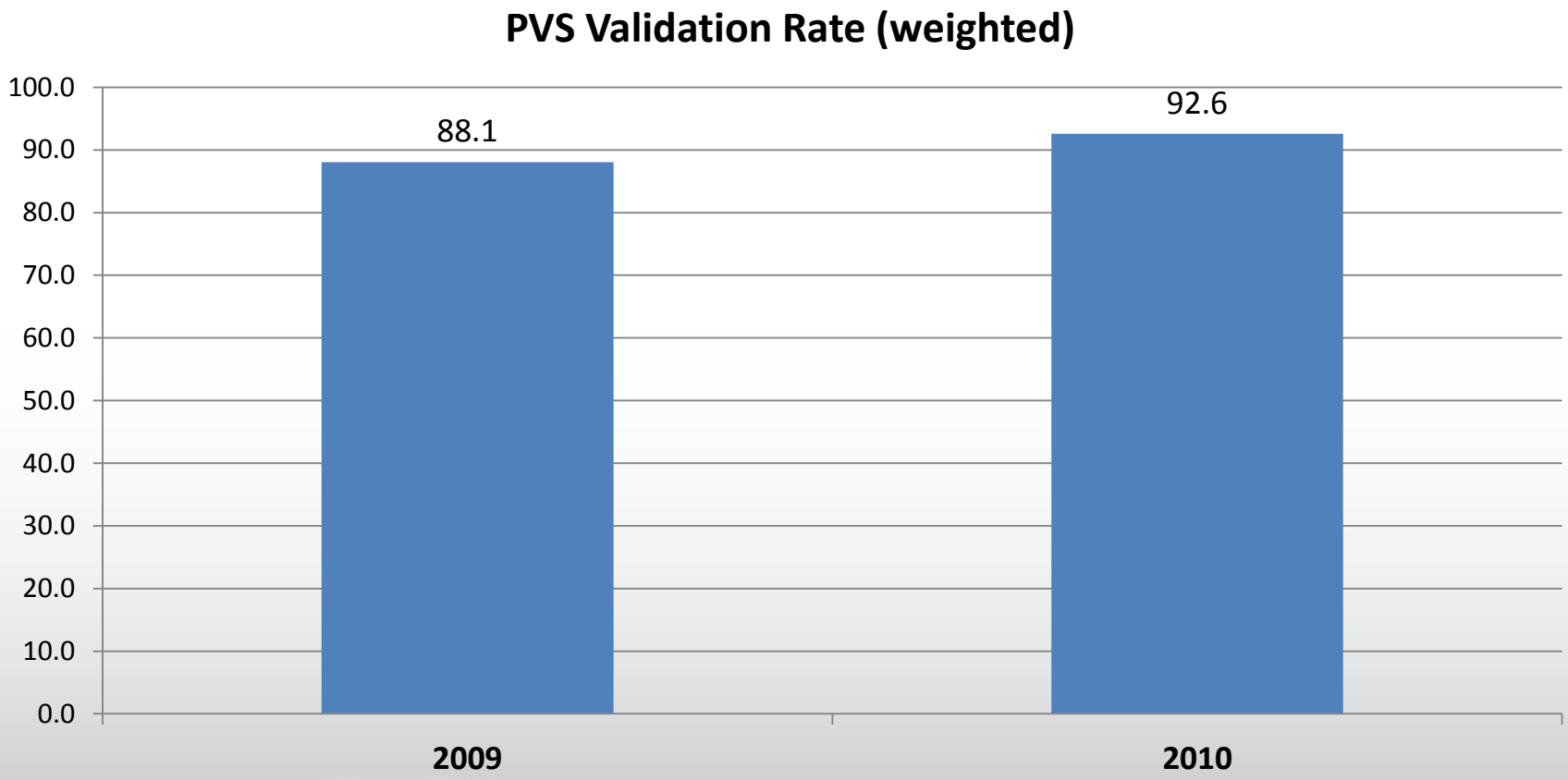
Expected Effects

Less likely to obtain a PIK:

- Insufficient or inaccurate person identifying info in incoming record
  - Issues w/ data collection or withholding due to language barriers, trust in govt., privacy preferences
  - Identifying info in incoming file & reference file more likely to differ
- Address info differs/not updated
  - Movers, rent vs. own, certain types of housing
- Record not in government reference files
  - Newborns, recent immigrant, very poor/unemployed/no govt. program recipient

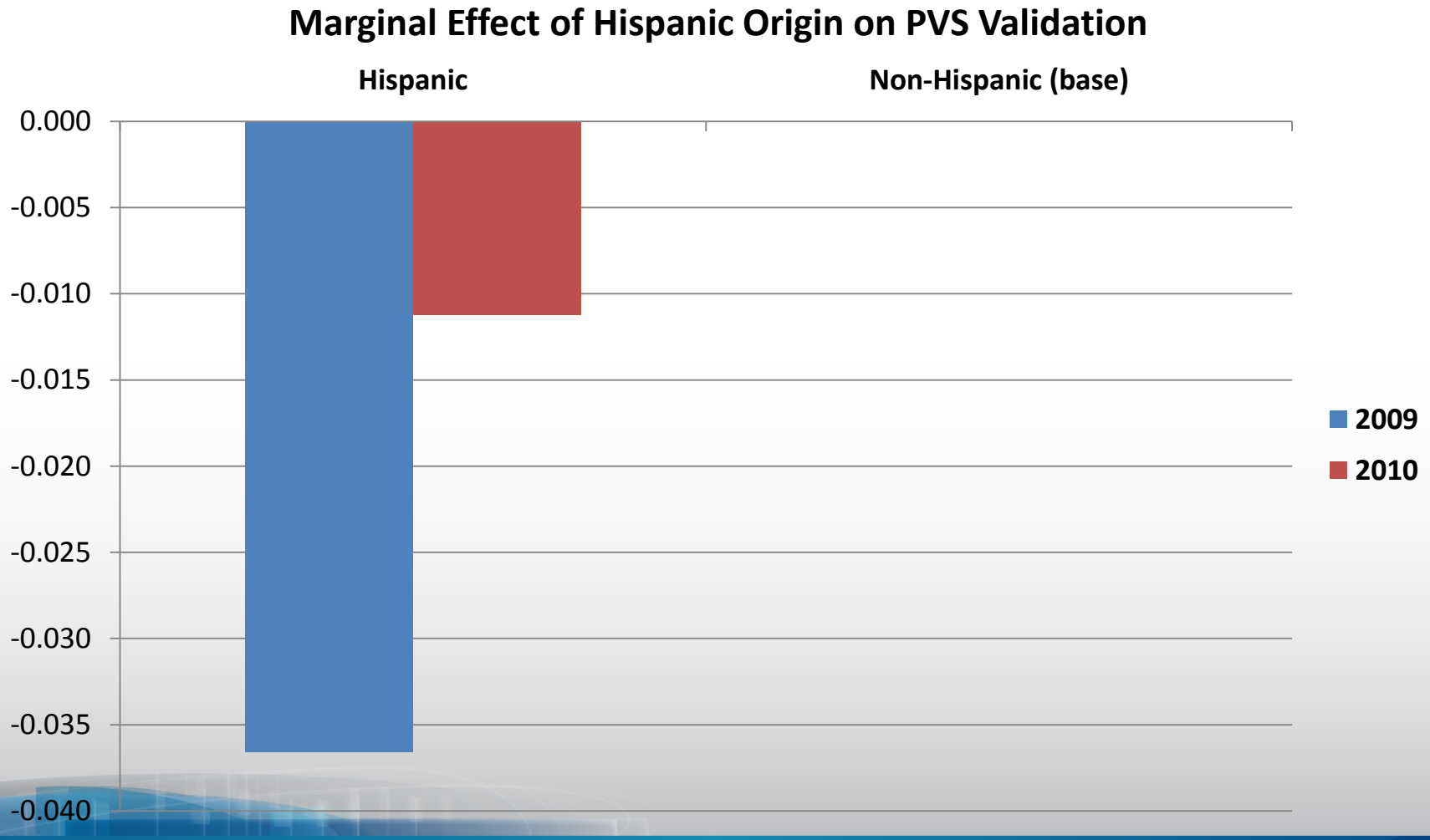
Results – Overall Validation Rates

Sources: 2009 & 2010 ACS



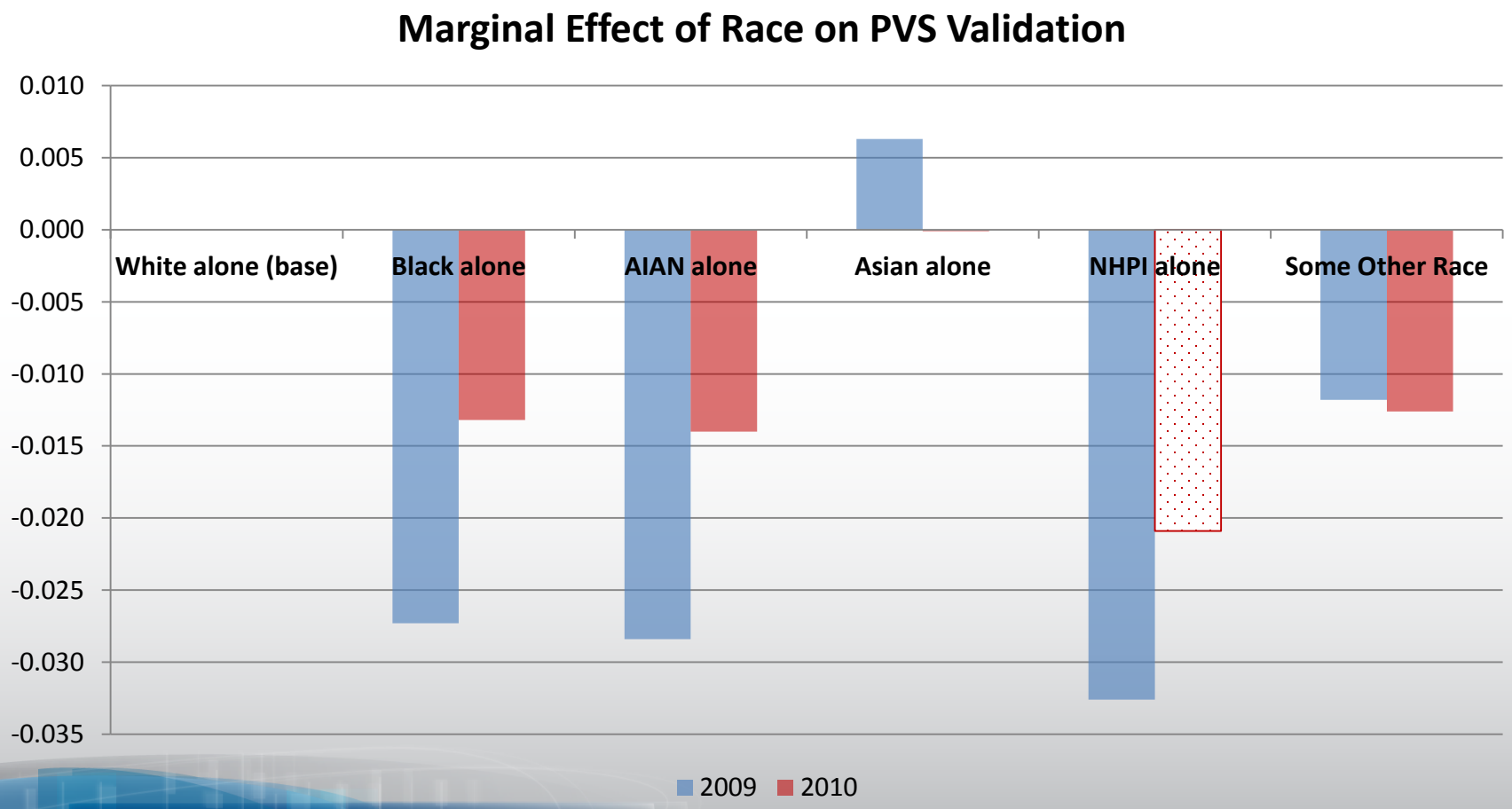
Probit Results

Sources: 2009 & 2010 ACS

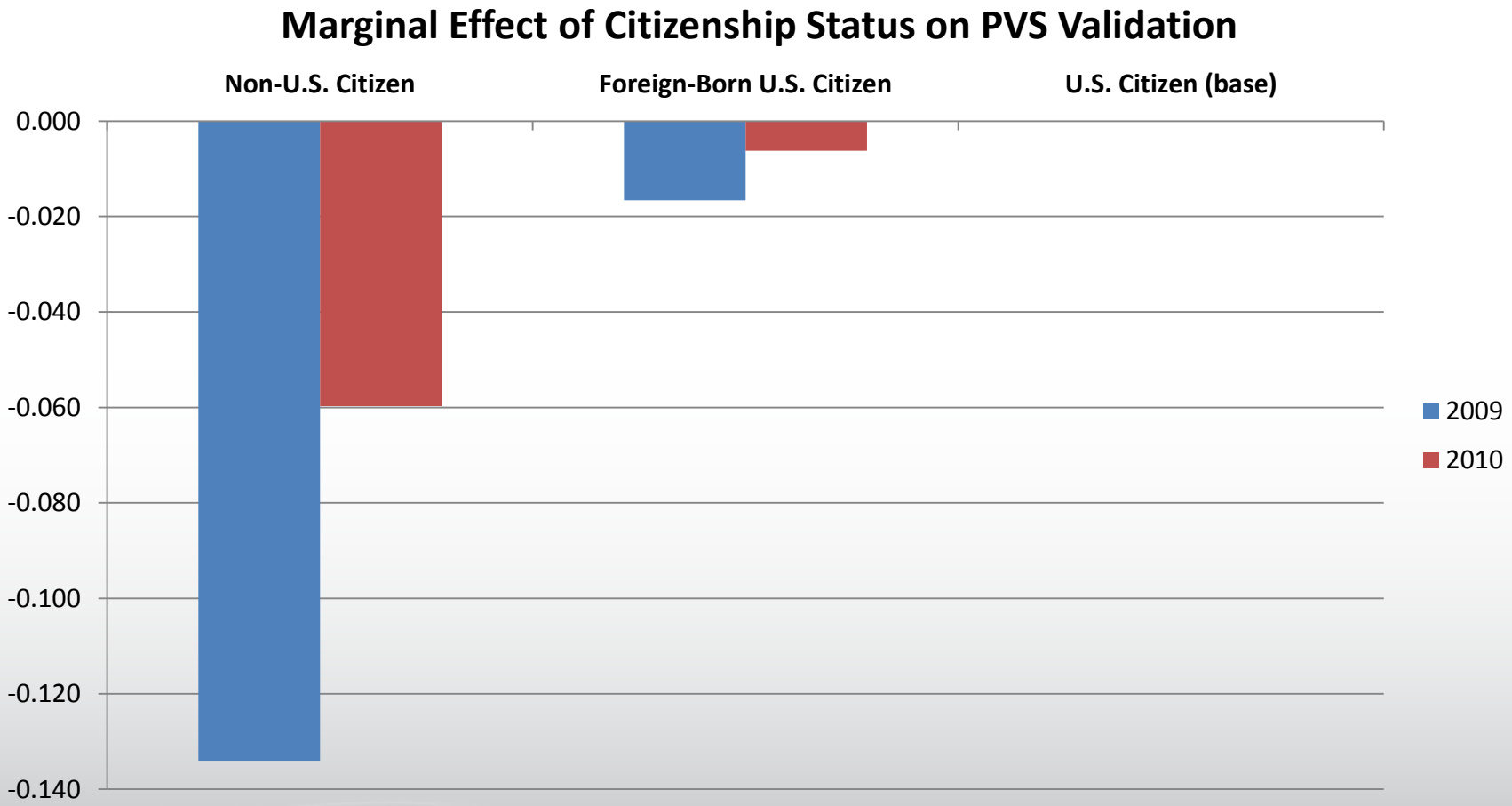


Probit Results

Note: Dotted bars indicate that change in marginal effect from 2009 to 2010 is not statistically significant.

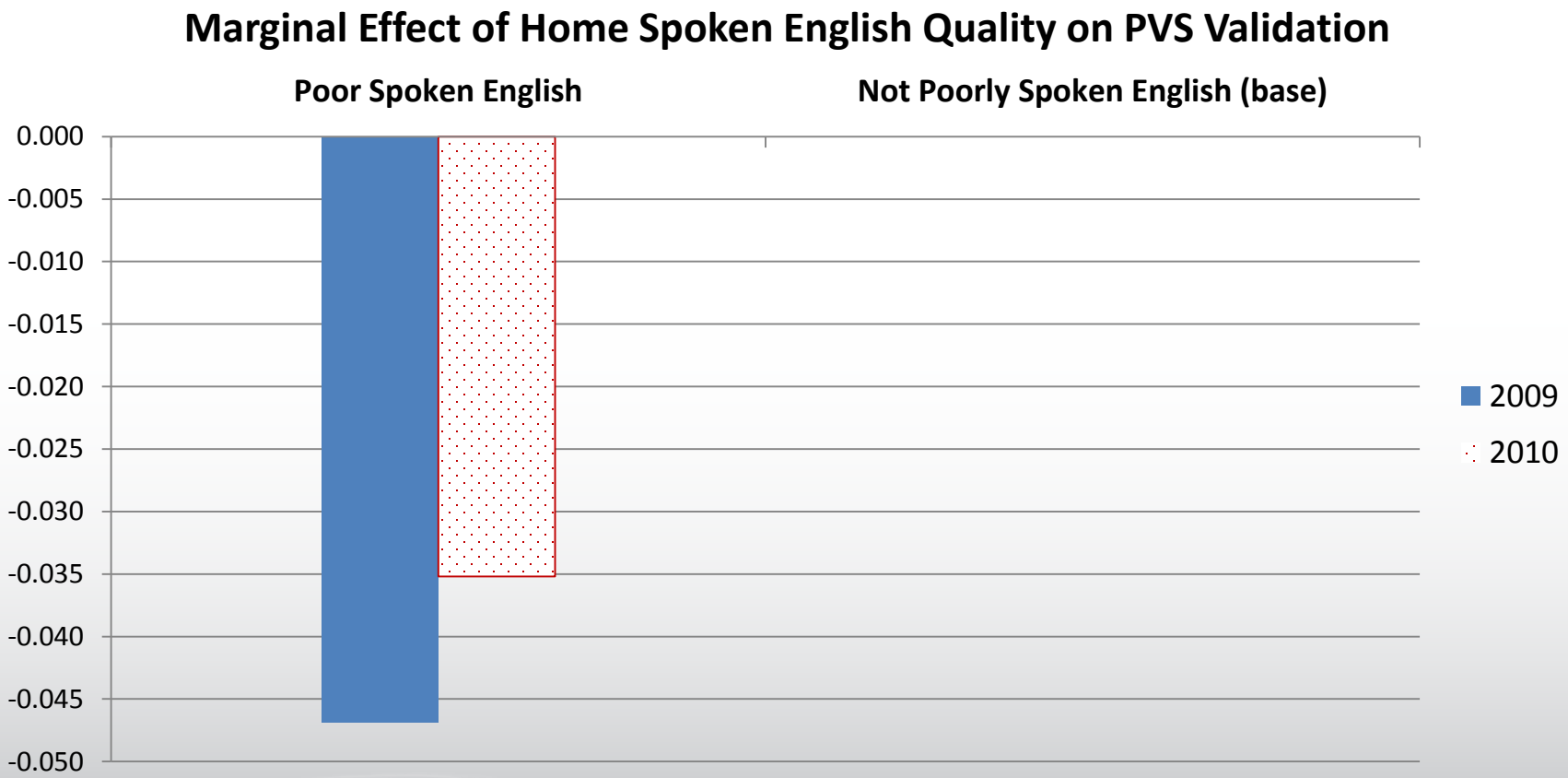


Probit Results



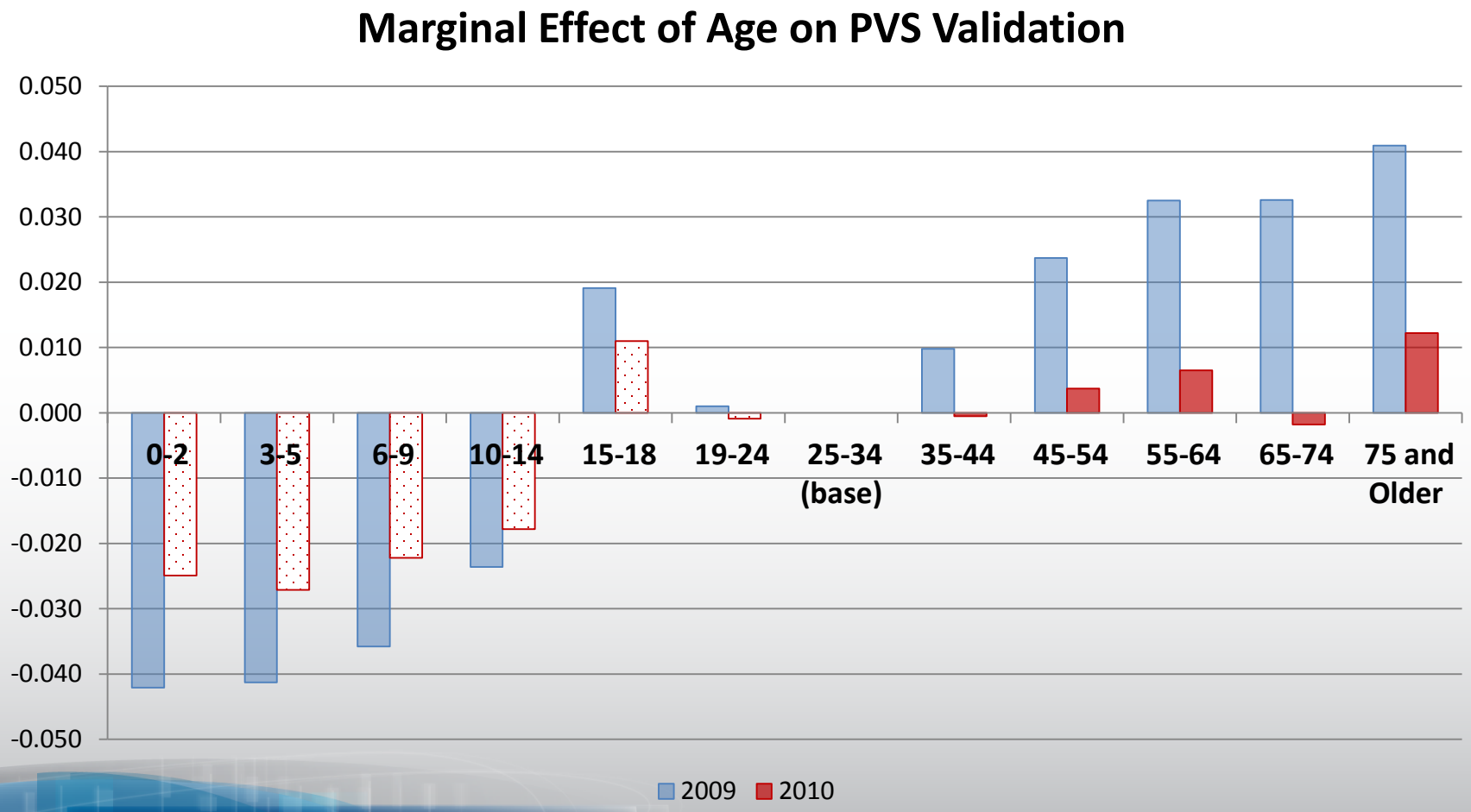
Probit Results

Note: Dotted bars indicate that change in marginal effect from 2009 to 2010 is not statistically significant.

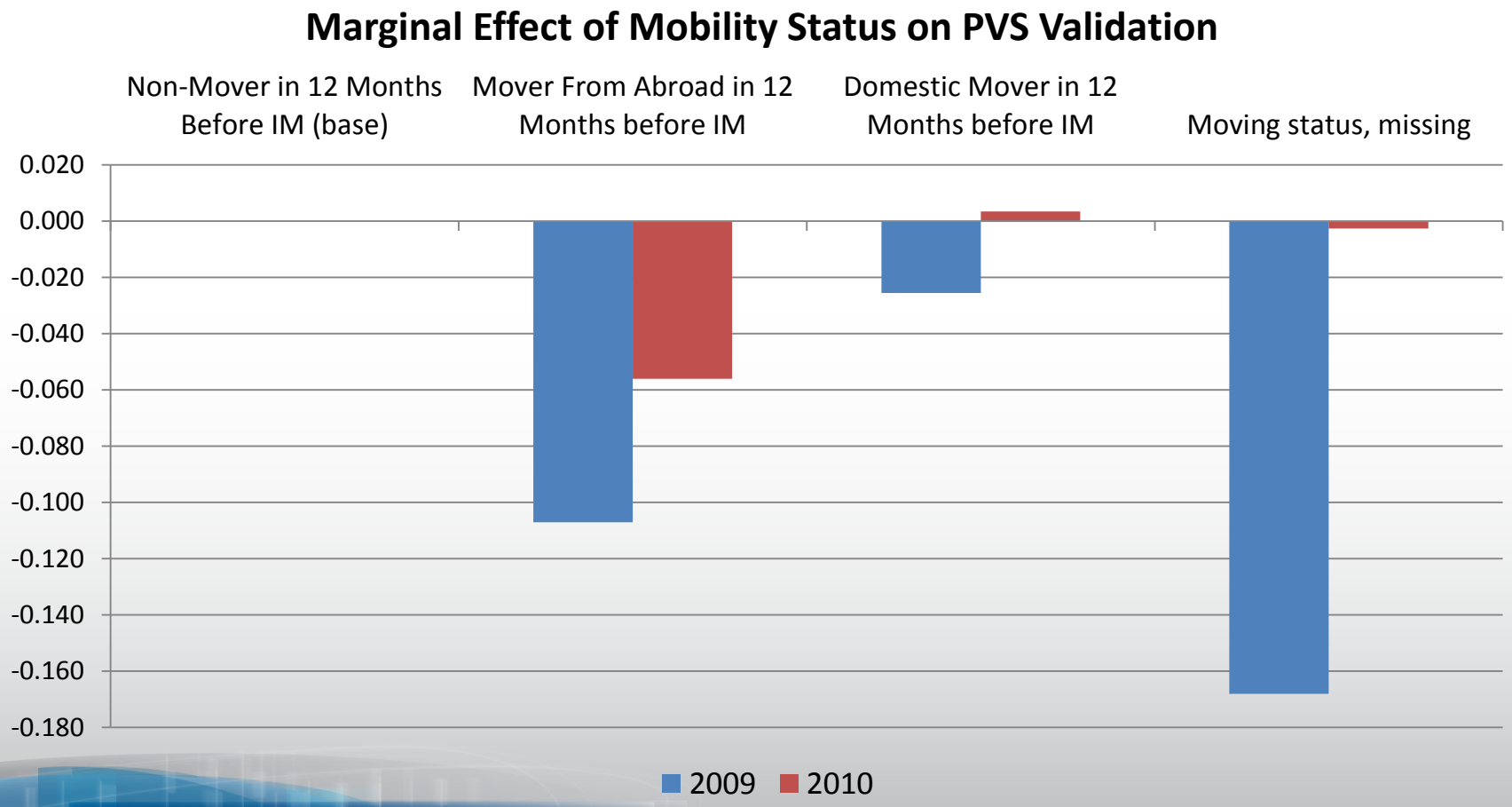


Probit Results

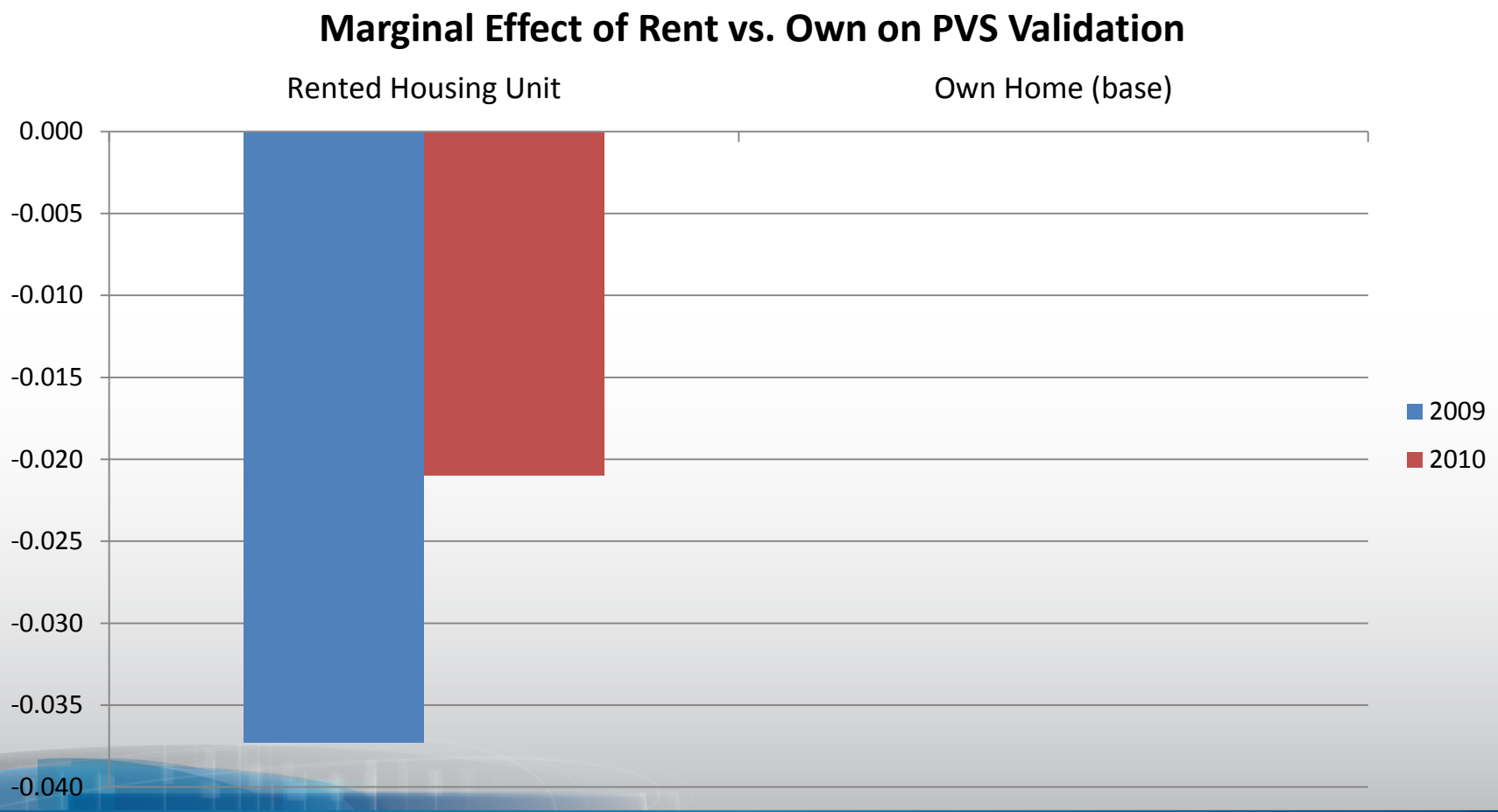
Note: Dotted bars indicate that change in marginal effect from 2009 to 2010 is not statistically significant.



Probit Results

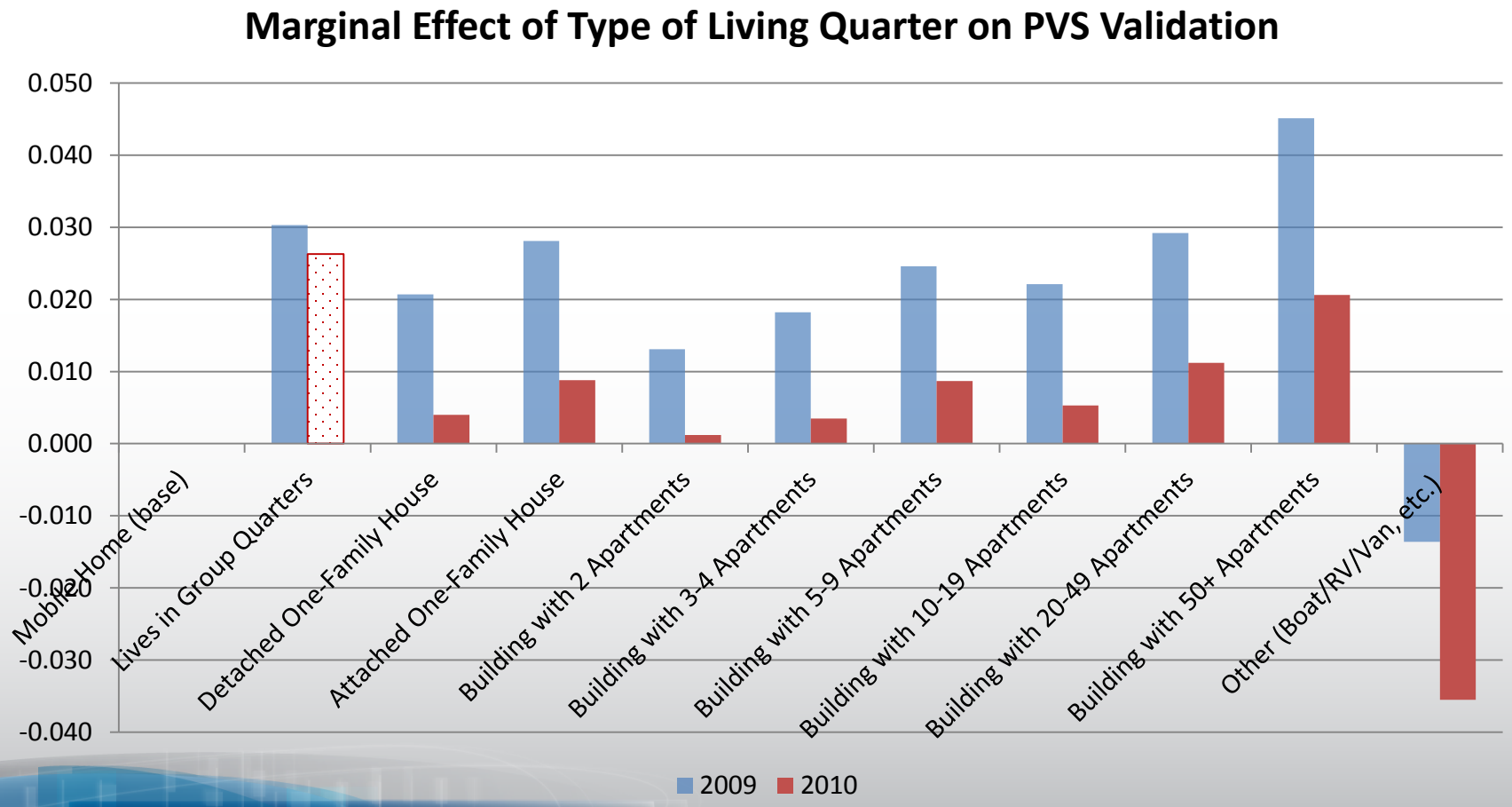


Probit Results

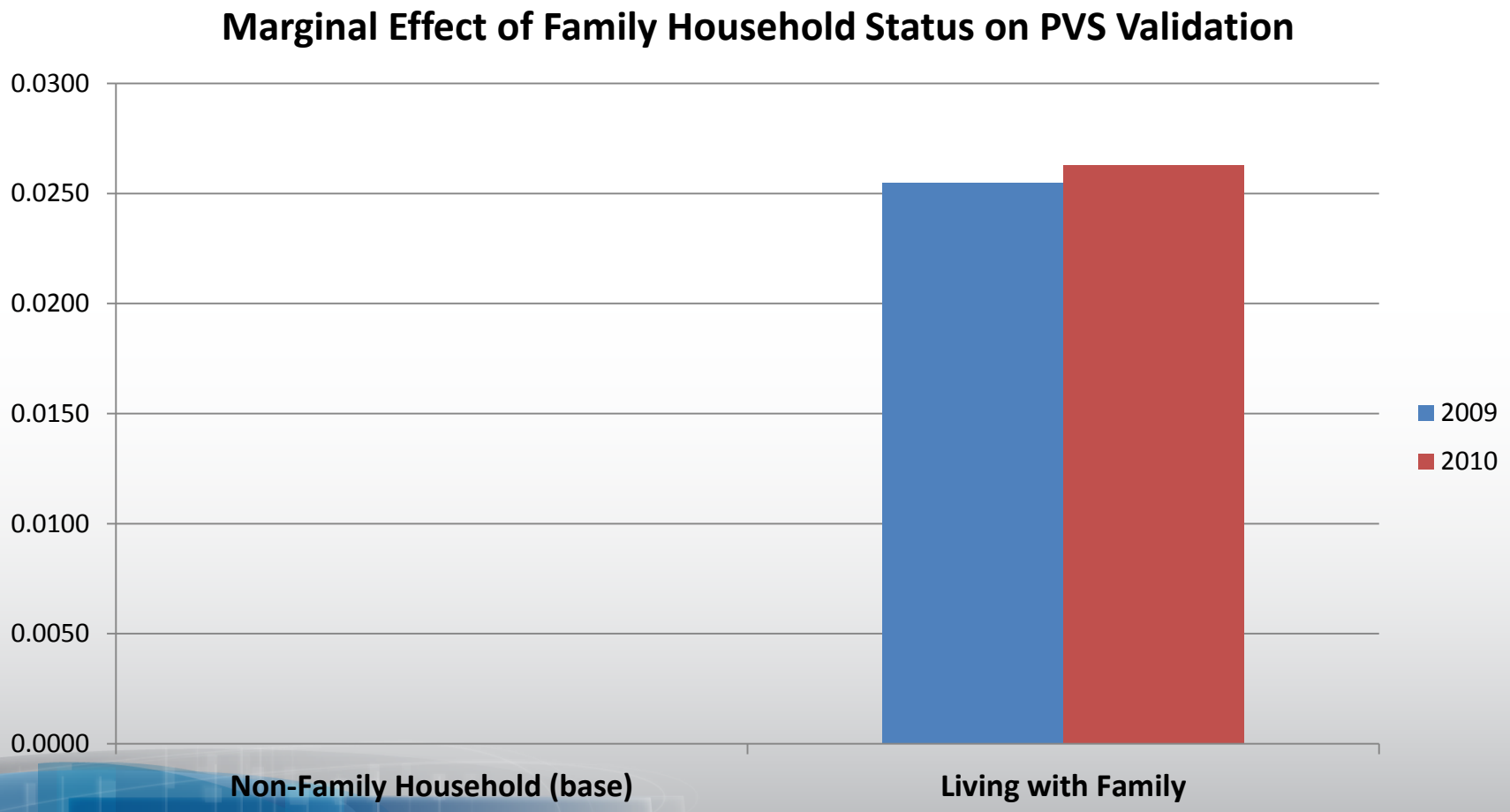


Probit Results

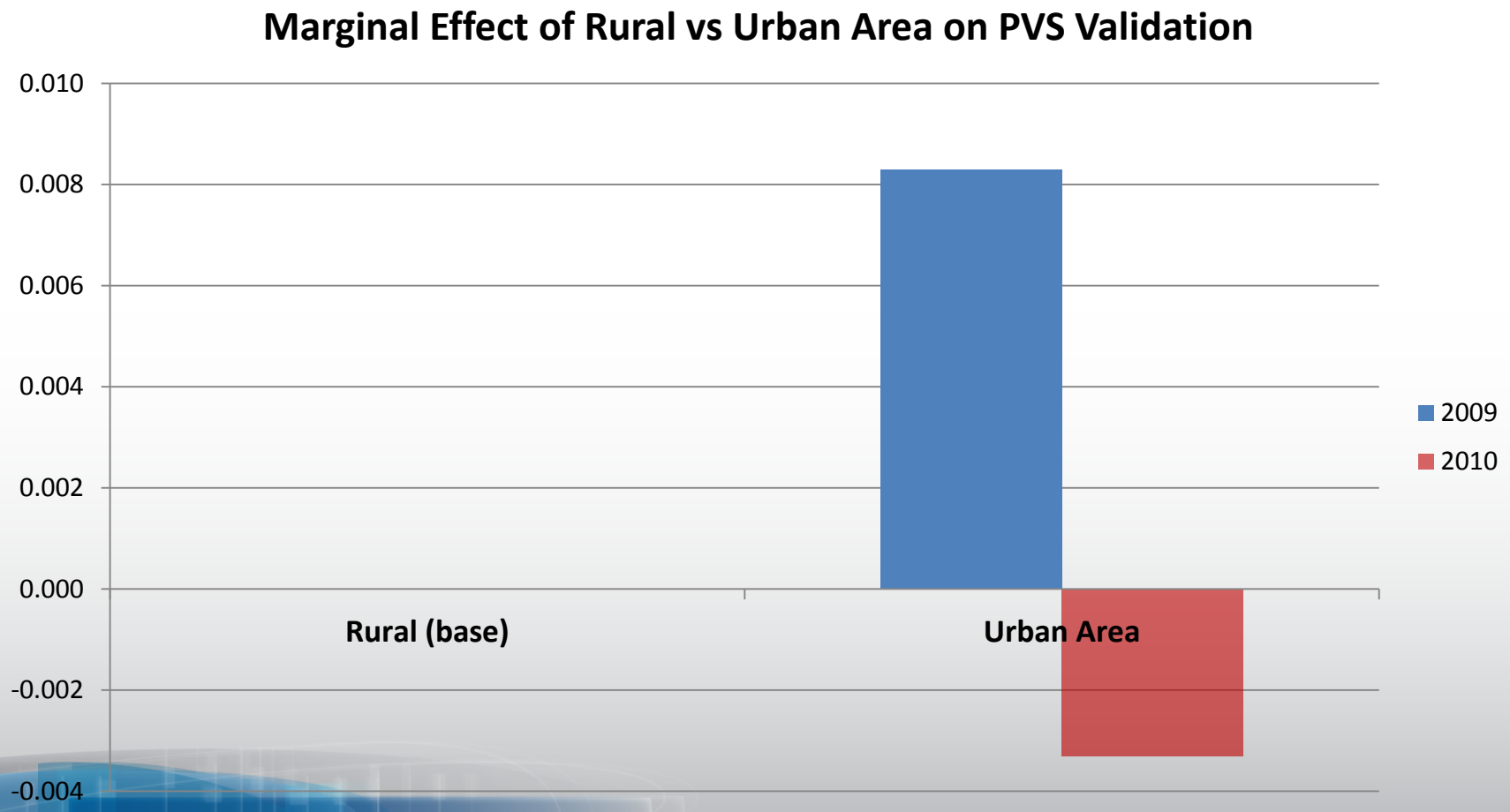
Note: Dotted bars indicate that change in marginal effect from 2009 to 2010 is not statistically significant.



Probit Results

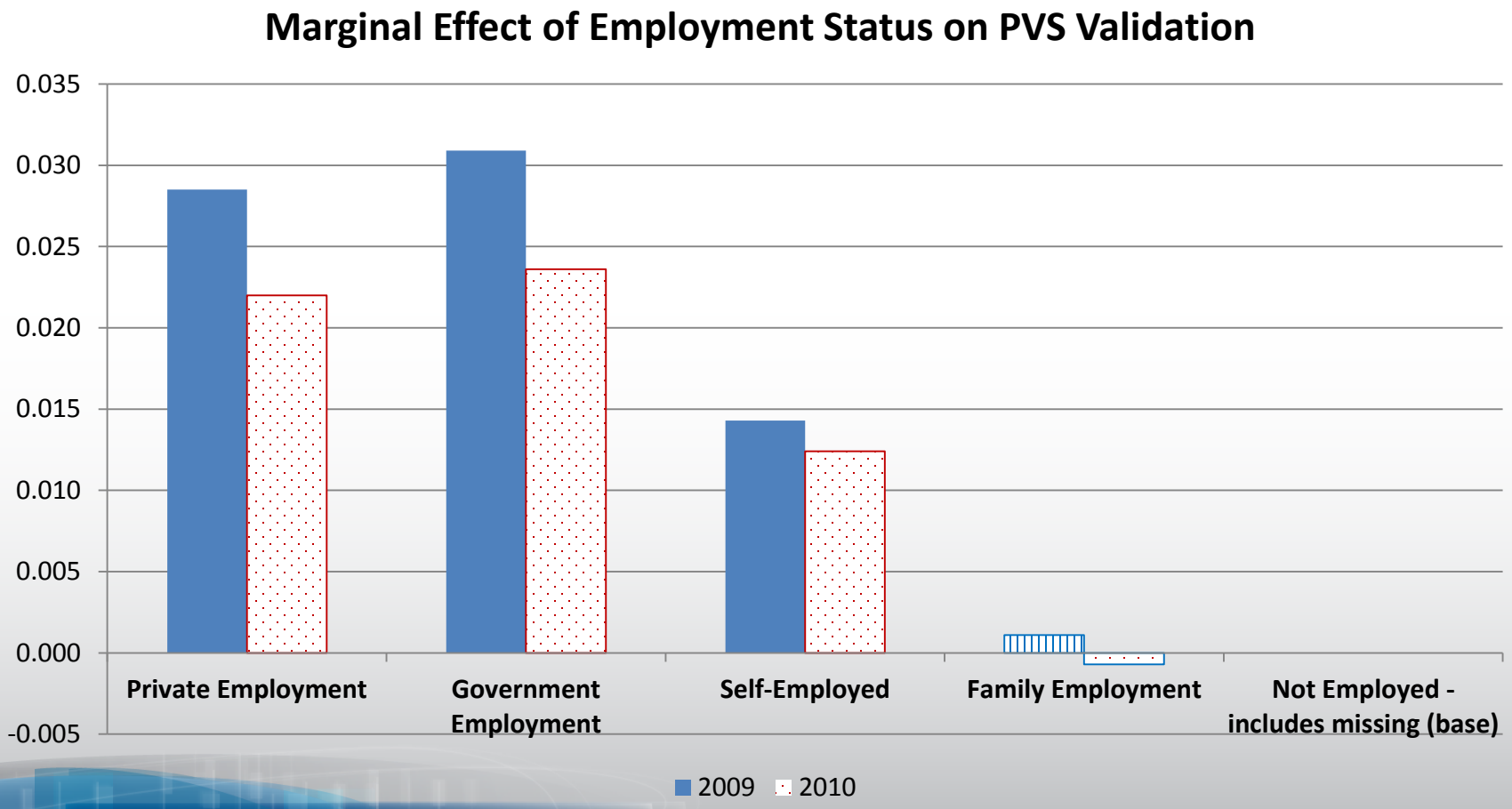


Probit Results

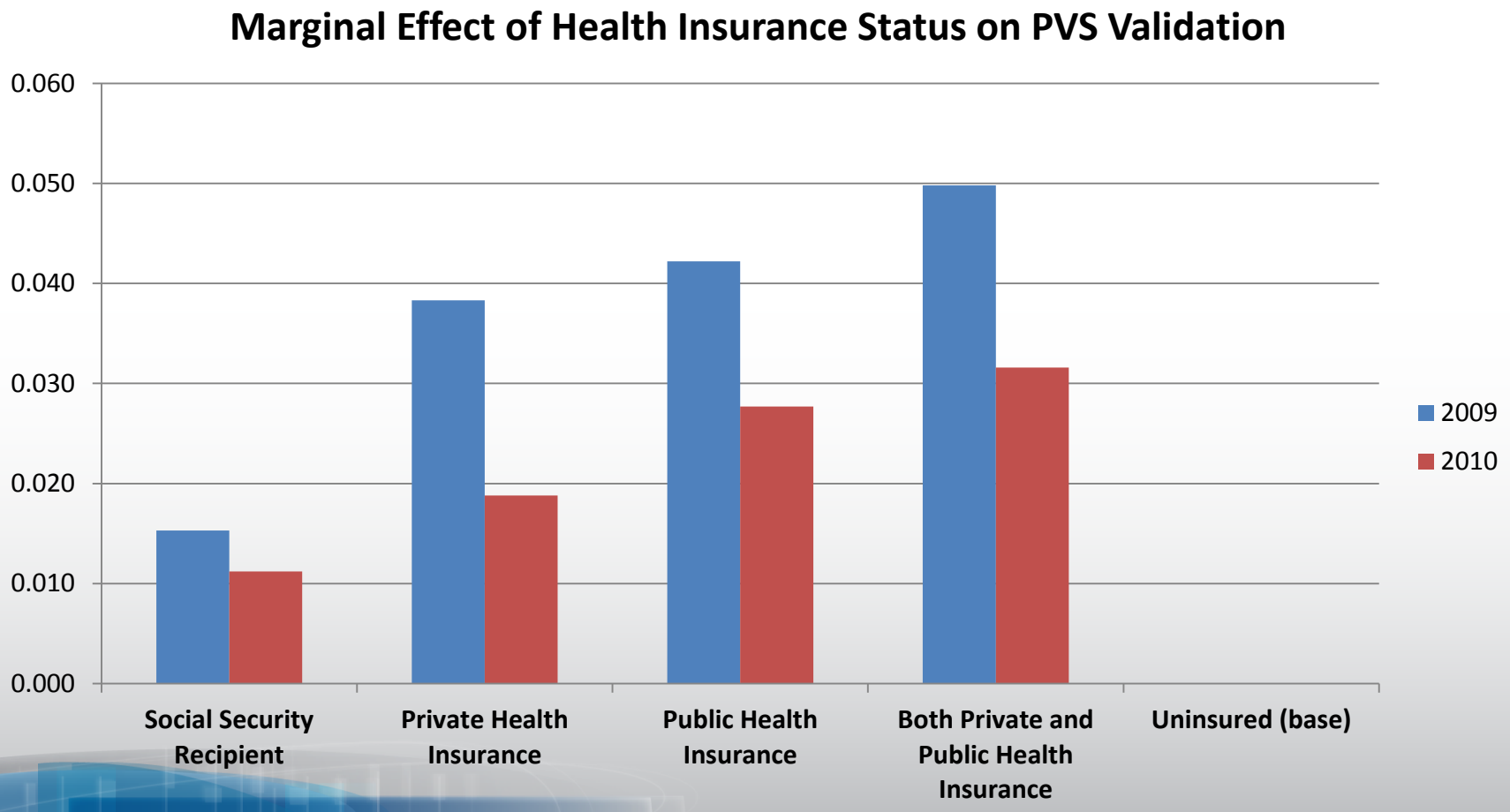


Probit Results

Note: Dotted bars indicate that change in marginal effect from 2009 to 2010 is not statistically significant.

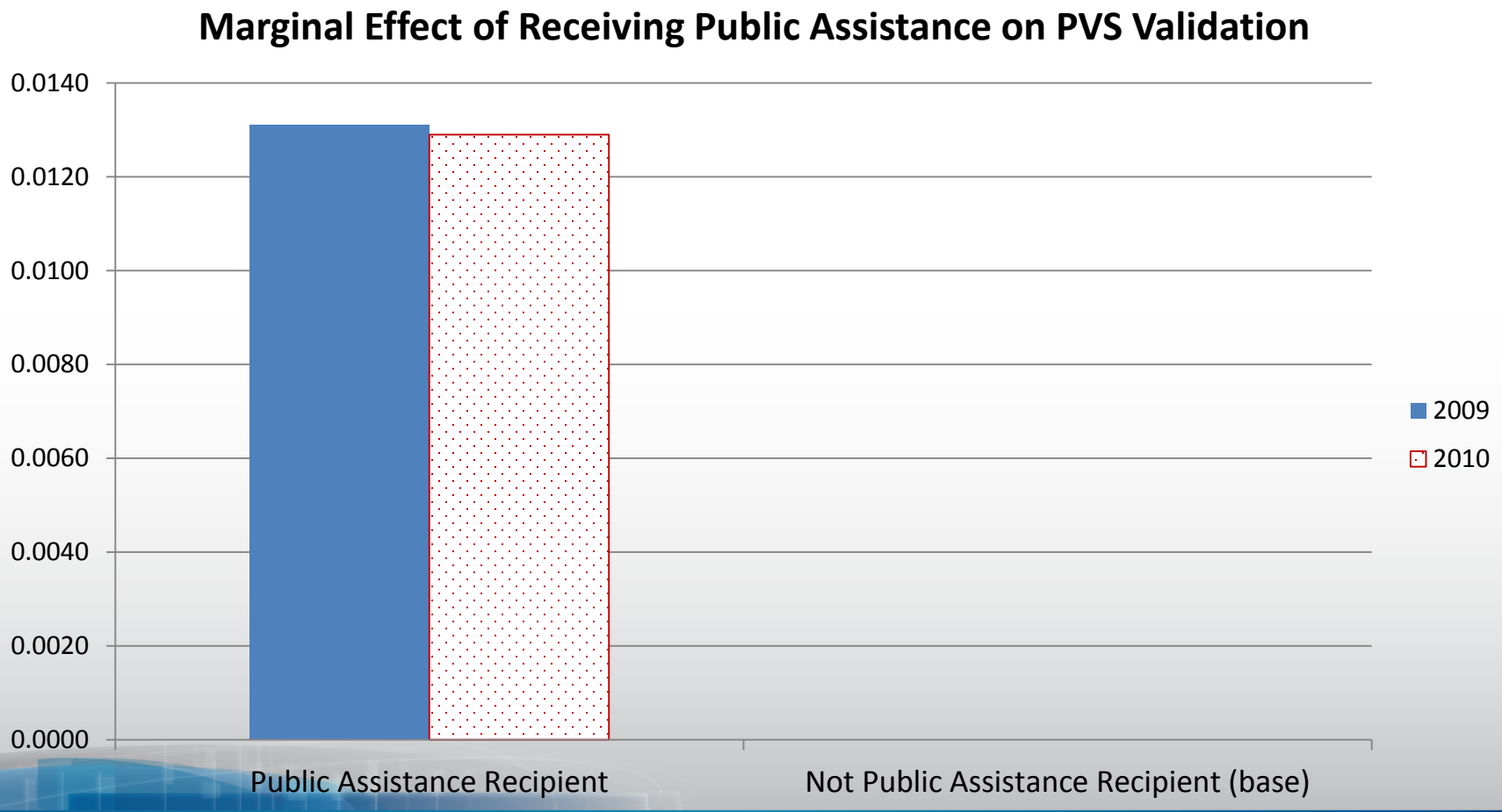


Probit Results

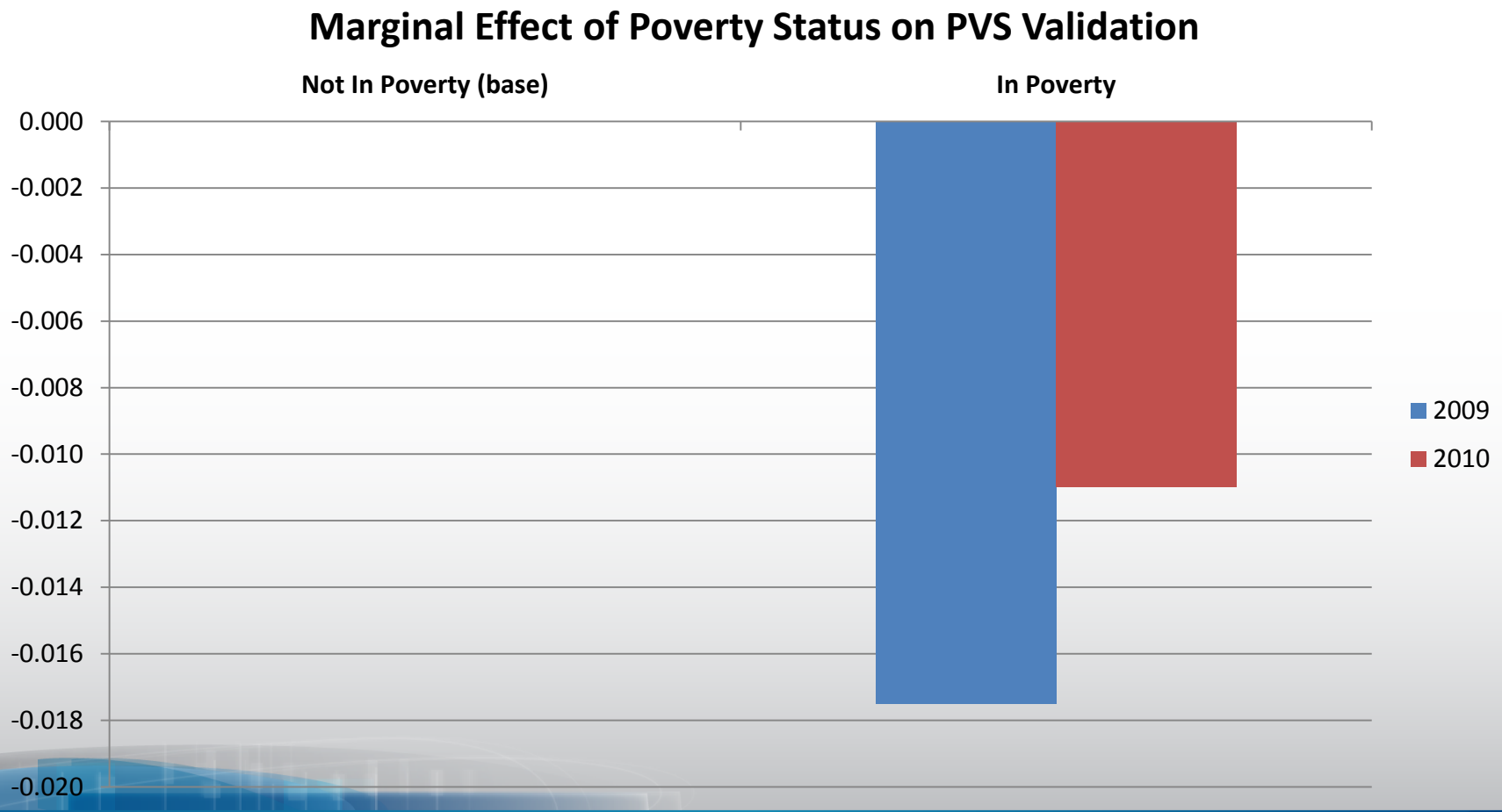


Probit Results

Note: Dotted bars indicate that change in marginal effect from 2009 to 2010 is not statistically significant.



Probit Results



Conclusions

- Mobile persons, those with lower income, unemployed, in process of integrating in economy/society, non-participants in government programs are less likely to be validated
  - Renters, movers, mobile homes
  - Low income, non-employed, most minorities, non-U.S. citizens, poor English
  - Non-participants of govt. program, uninsured, non-military
- Researchers may wish to reweight observations based on validation propensity

Conclusions

- Changes to PVS system
  - Increased overall validation rate by 4.5 percentage points
  - Reduced validation differences across most groups from 2009 to 2010

- Record linkage research can lead to higher PIK assignment rates and less bias

Thank you!

adela.luque@census.gov