

On the Quality of Ancillary Data Available for Address-Based Sampling

Charles DiSogra, Knowledge Networks
J. Michael Dennis, Knowledge Networks
Mansour Fahimi, Marketing Systems Group

Presented at the Joint Statistical Meetings
August 3, 2010
Vancouver, BC

KnowledgePanel®

Probability-based, nationally representative panel of US adult population

Includes:

- Households with no Internet access, KN provides laptop computer, free ISP
- Spanish-language dominant households

Extensive profile data collected and maintained on members

All surveys administered by Web mode



ABS Mail Recruitment

Since 2009, Knowledge Networks panel members recruited via mail using an address-based sample frame (ABS)

Sample provided by Marketing Systems Group (MSG)

- U.S. Postal Service Computerized Delivery Sequence File (CDSF)
 - >97% coverage of physical addresses
 - Frequent updates including status of addresses, such as seasonal homes, vacation homes, vacant houses, etc.

- MSG can also provide:
 - telephone number match for ABS sample addresses
 - latitude-longitude location
 - ancillary demographic data for purposes of non-response analysis, custom mail targeting, etc.

Ancillary Data Used by KN Provided by MSG

- Household level
 - ✓ Telephone number (landline, match rate 60%+, used for non-response follow-up calls)
 - ✓ Latitude, longitude
 - ✓ Number of adults
 - ✓ Presence of children (yes, no)
 - ✓ Home ownership (own, rent)
 - ✓ Household income (12 levels recoded to <25k, 25-49k, 50-74k, 75k+)

- Person level
 - ✓ Marital status (married, single)
 - ✓ Education of head of HH (<HS, HS, Some College, BA, Higher)
 - ✓ Age of head of HH (used in KN's pilot sample only)
 - ✓ Race/ethnicity (33 codes recoded to White, African American, Hispanic, Other)

- MSG sources: Several databases, including infoUSA, Experian, Acxiom

- Unavailable data can be high, range 5-27%
 - Highest availability rates for income and presence of children; lowest for race/ethnicity & education
 - Unavailability rates include in the sample the vacant, seasonal, PO Box and Rural Route addresses

Example: Ancillary Data Response Analysis

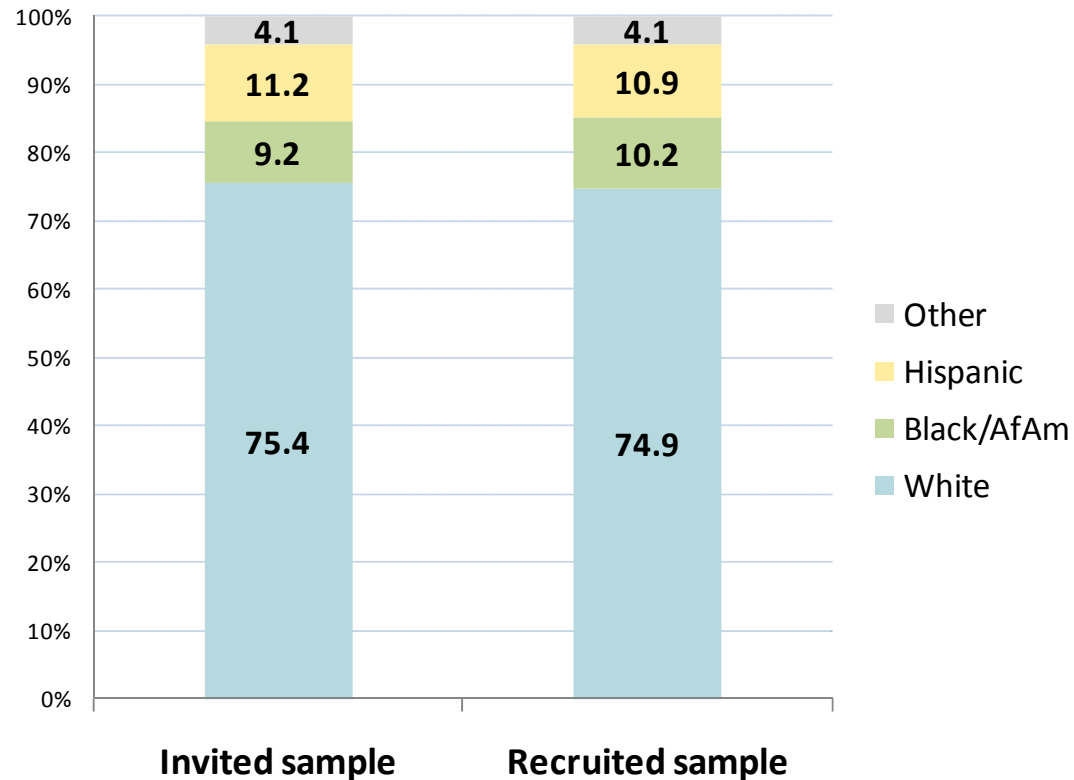
Purpose: Non-response bias test

Analysis: Compare ABS invited sample to the subset of actual KN recruited sample using the ancillary data

- April 2009 mailing of recruitment invitations
- “Invited”: 40,000 nation-wide ABS sample addresses
- “Recruited” subset: 4,000 KN recruited households

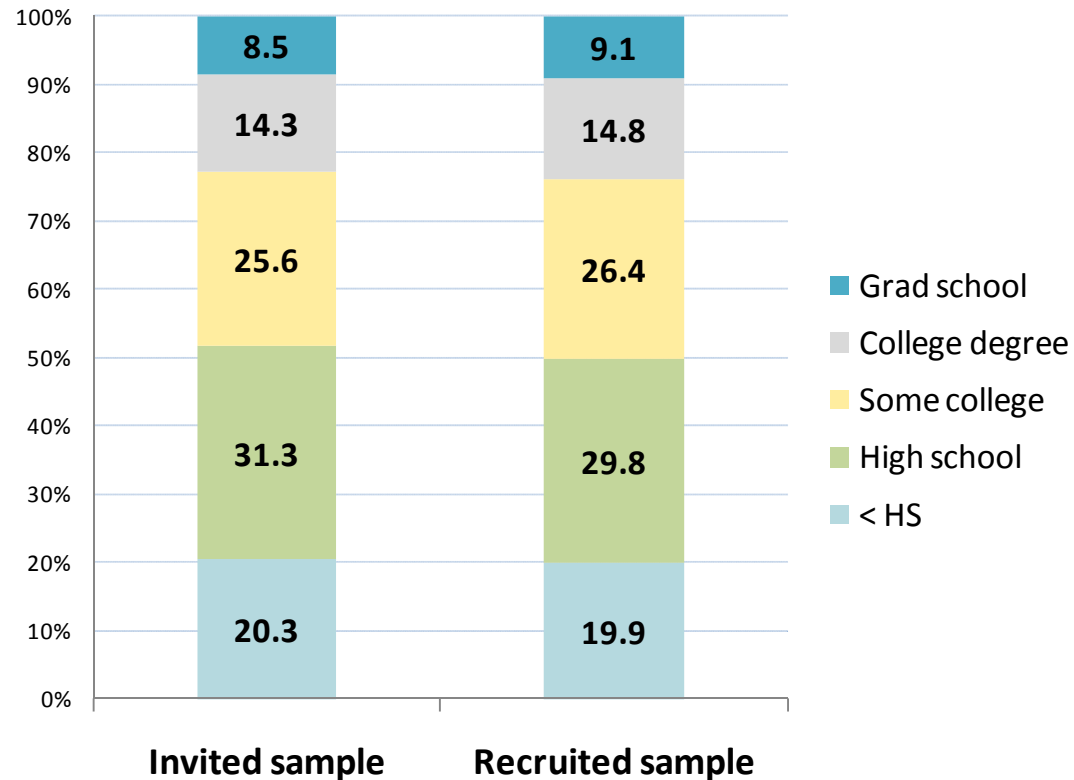
We found excellent alignment of the approx. 4,000 recruited addresses with all invited addresses using MSG’s descriptive ancillary data

Race/Ethnicity Response Comparison



40,000 invited, ~4,000 recruited; April 2009 data

Education Level Response Comparison



40,000 invited, ~4,000 recruited; April 2009 data

Household Income Response Comparison



40,000 invited, ~4,000 recruited; April 2009 data

Ancillary Data Research Focus

How accurate are these ancillary data as “predictors” of the sample household?

Matching Analysis

We compared the two household data sources:

- Predicted ancillary data
matched to
- Actual survey information collected from KN panelists recruited from these households for whom we have predicted ancillary data

Sample Size

- 10,000+ KN panel households recruited 2008-early 2010

HOH Issue: Measuring head of household is complicated by:

- Fact that ancillary data sources use different definitions for HOH;
- Declining cultural relevance of the concept;
- Varying household-level interpretations of the concept

KN Groups Compared to Predicted Ancillary Data

All Recruited KN Panelists:

- Universe of recruited and profiled KN adult panel members n=10,000+ (includes multiple per household)

Subset of Recruits: Primary KN Panel Respondent:

- First adult listed on paper recruitment form or entered first online or person we talked to during non-response phone recruitment (who supplies information for self and other household members)

Subset of Recruits: KN Panel Head of household:

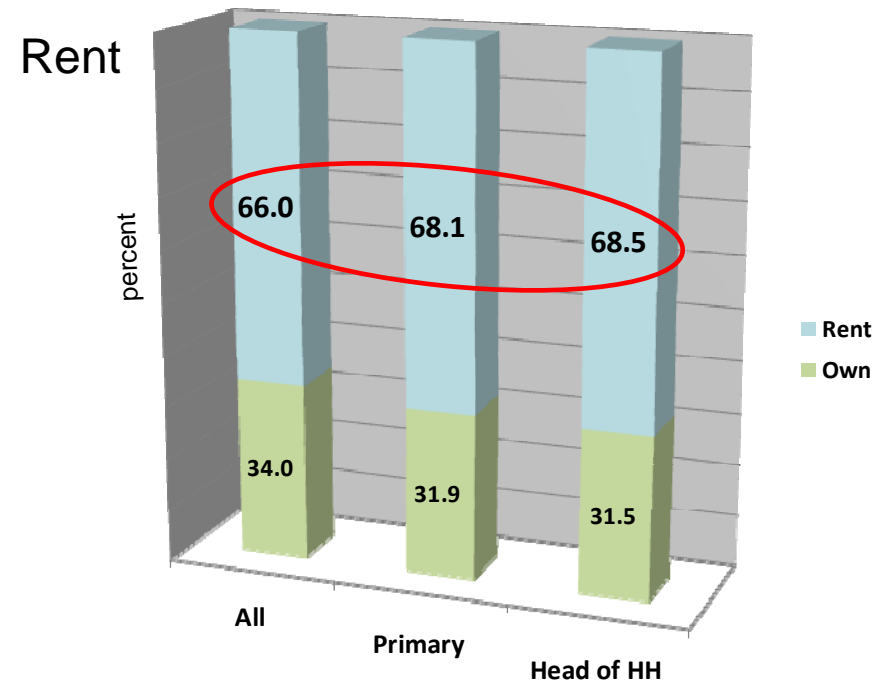
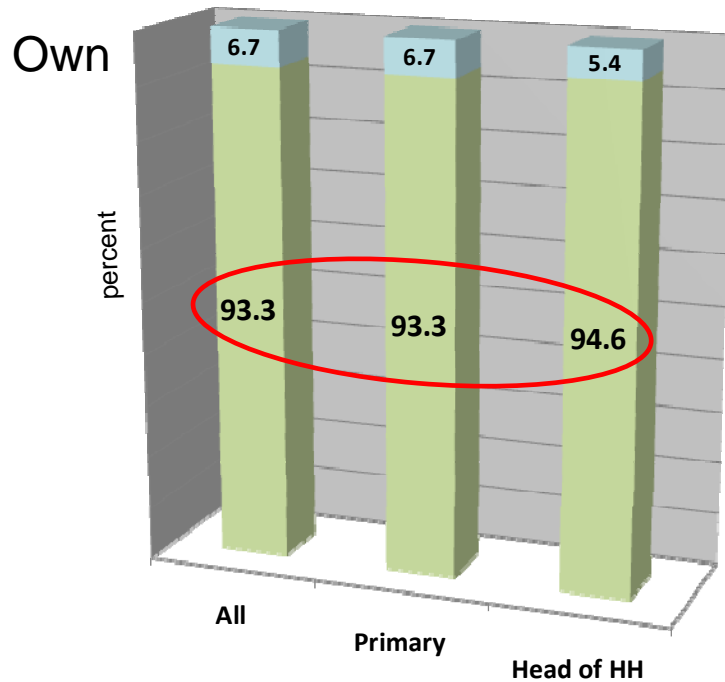
- House/apartment is in adult panelists' name, that adult is considered head of HH
- If there is more than one such panelist in the household, the oldest male is selected as head of HH

Correlations: Predicted Ancillary Data with KN Panel Member Data

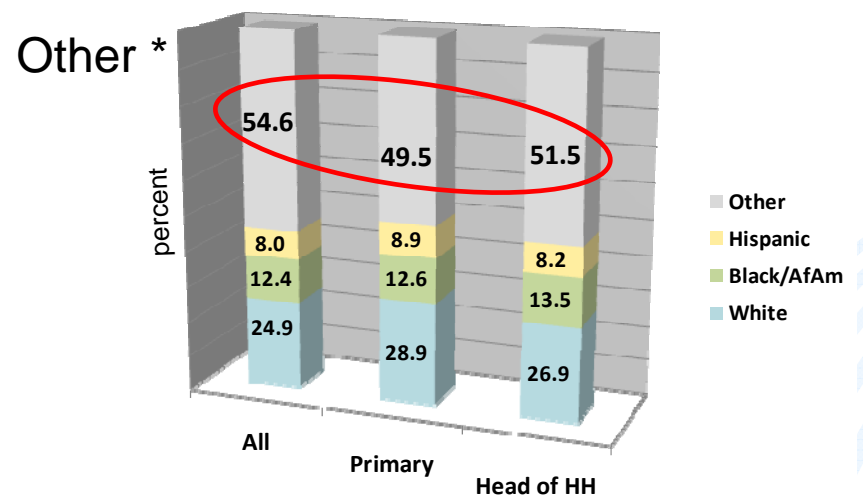
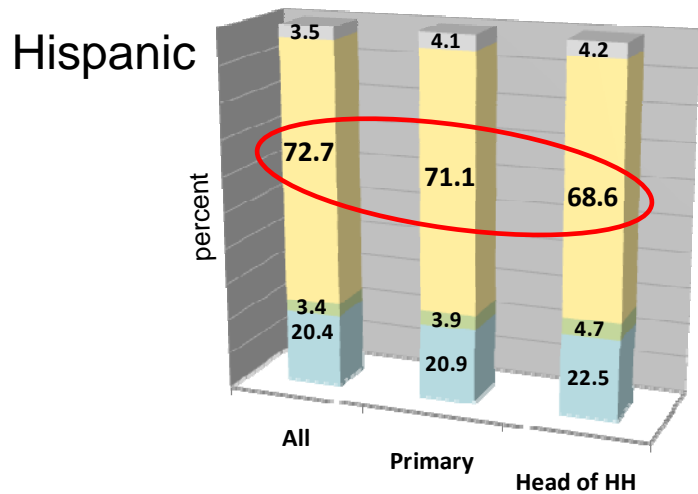
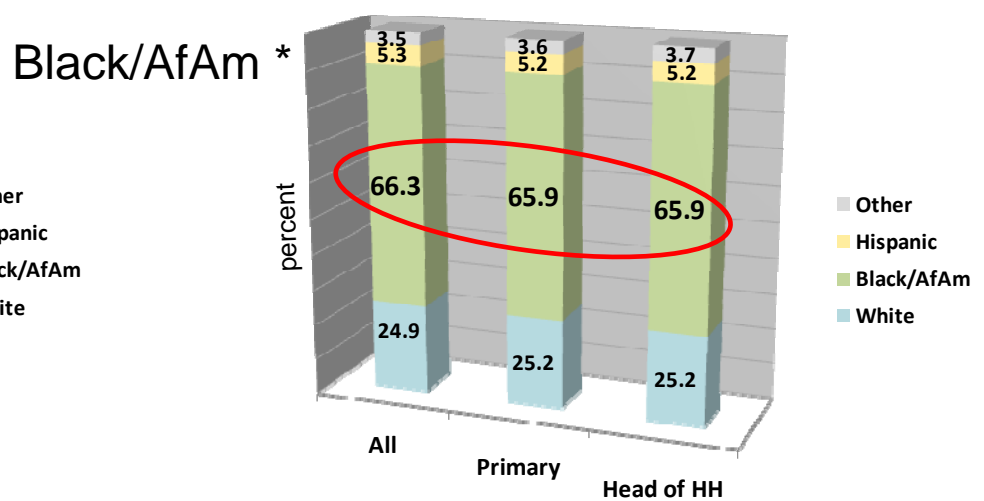
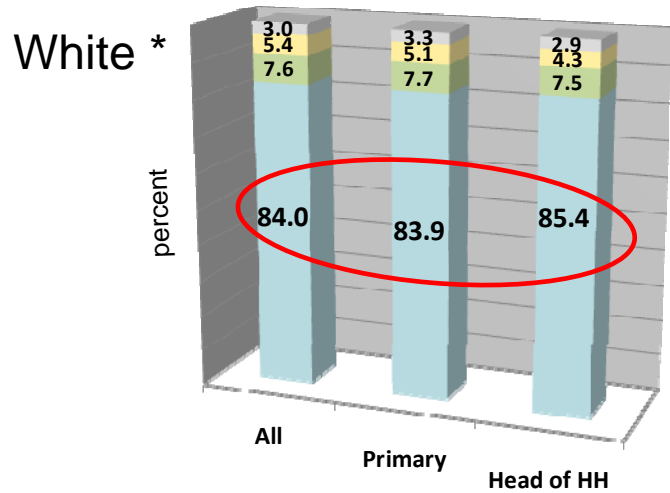
Ancillary variables (ordered by Head of HH rank)	All			Primary			Head of HH			Strong ↓ Weak
	rank	r^*	n	rank	r^*	n	rank	r^*	n	
Home ownership	1	0.634	10,480	1	0.652	7,727	1	0.675	7,045	
Age Head of HH (pilot data)	3	0.593	437	2	0.625	391	2	0.665	366	
Race/ethnicity	2	0.619	8,880	3	0.608	6,509	3	0.608	5,894	
Marital status	4	0.467	10,480	4	0.502	7,727	4	0.546	7,045	
Household income	5	0.445	11,162	5	0.456	8,234	5	0.470	7,484	
Children in household	7	0.357	11,537	7	0.367	8,496	6	0.386	7,716	
Education of Head of HH	6	0.365	9,302	6	0.379	6,839	7	0.385	6,198	
Number of adults	8	0.261	10,480	8	0.281	7,727	8	0.302	7,045	

* All correlations are significant at $p < .0001$

Predicted Home Ownership vs. KN Actual % by Group



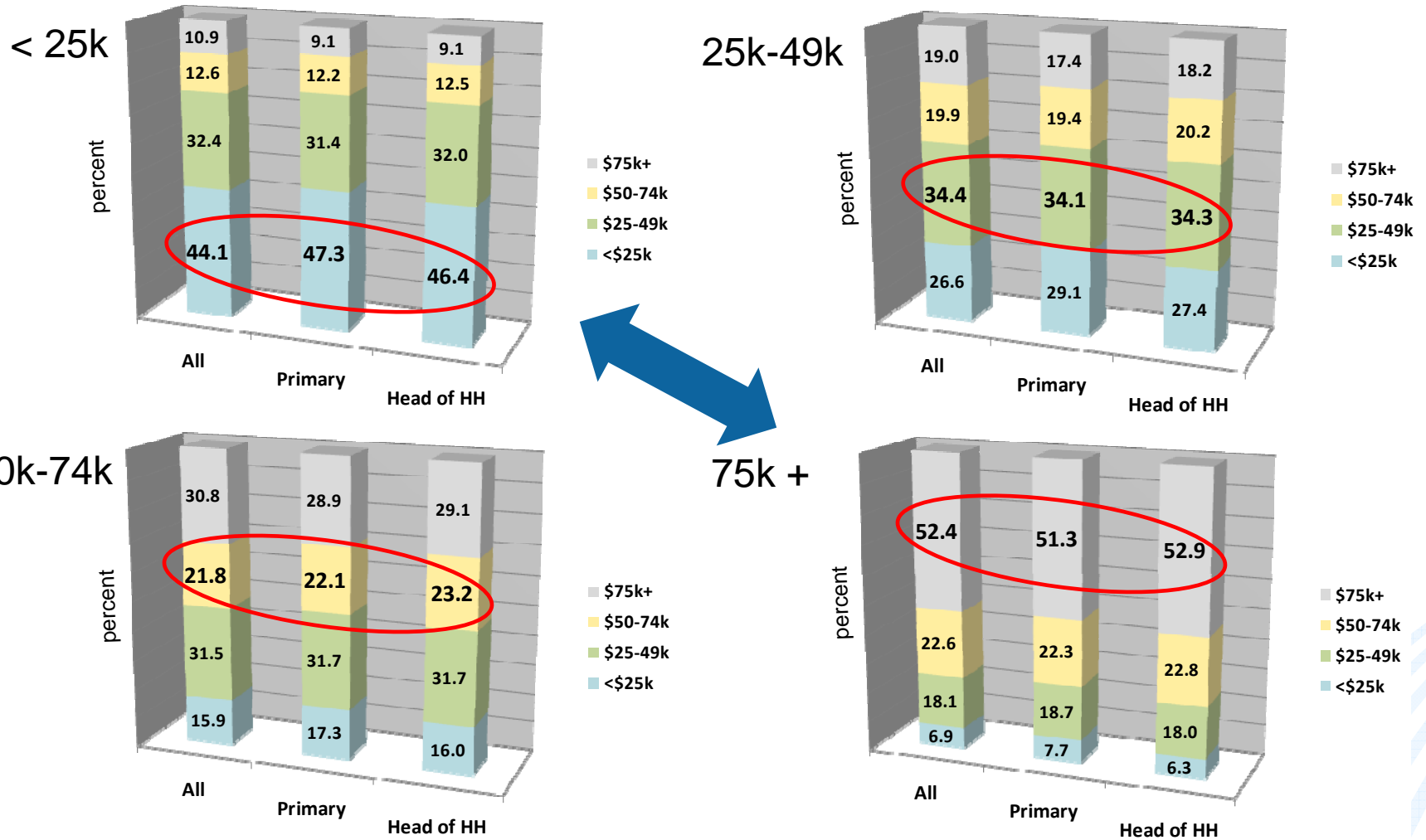
Predicted Race/Ethnicity vs. KN Actual % by Group



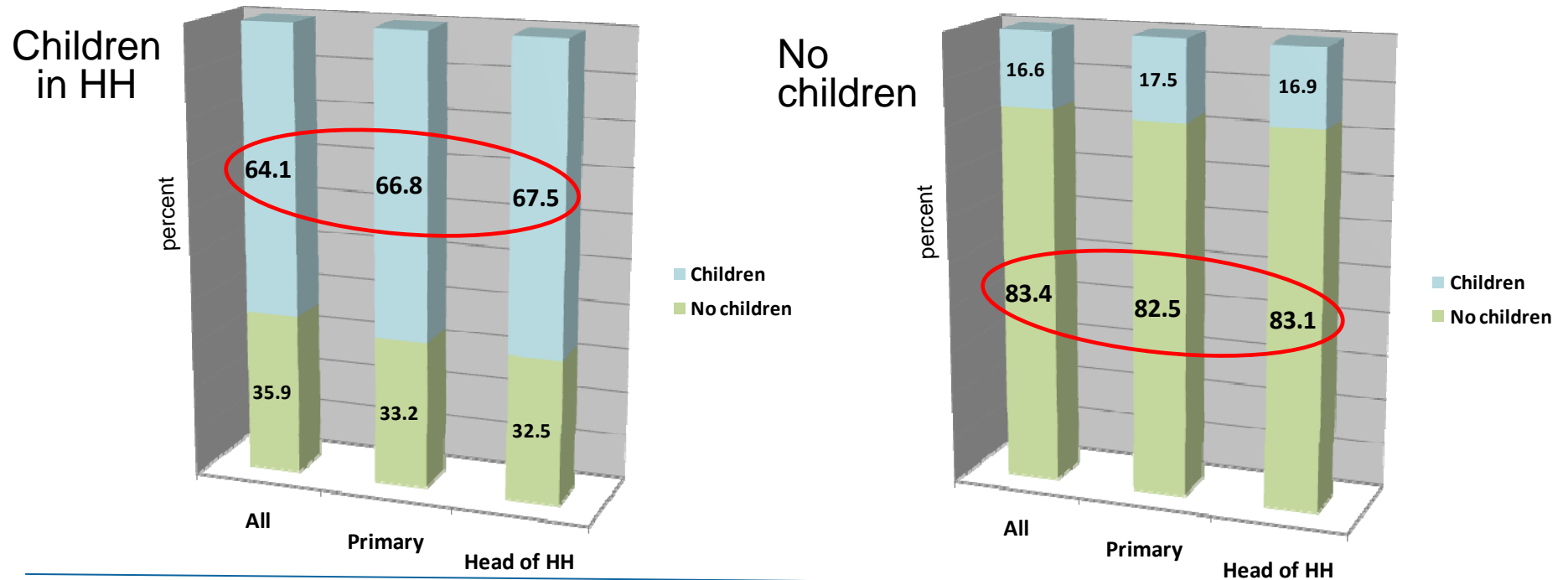
Predicted Marital Status vs. KN Actual % by Group



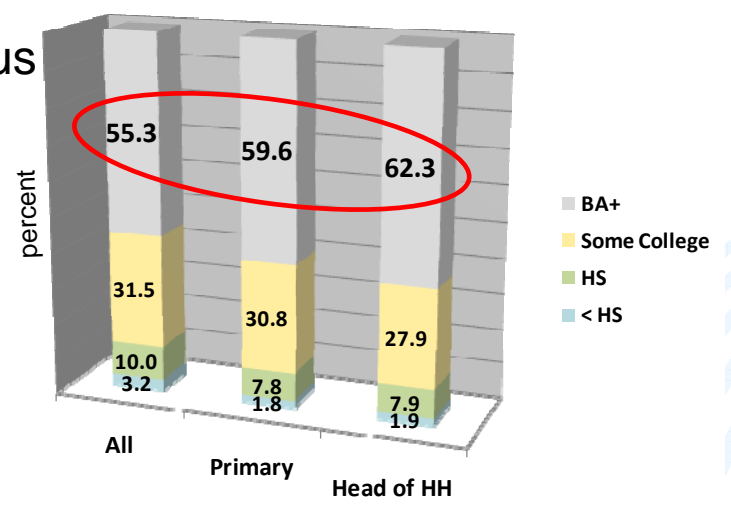
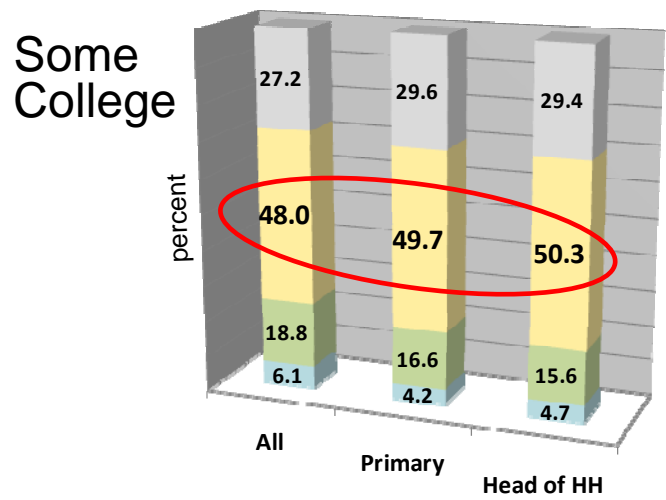
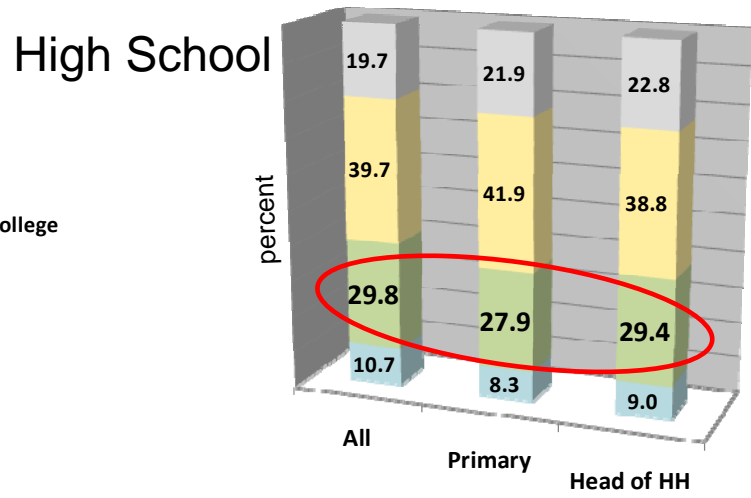
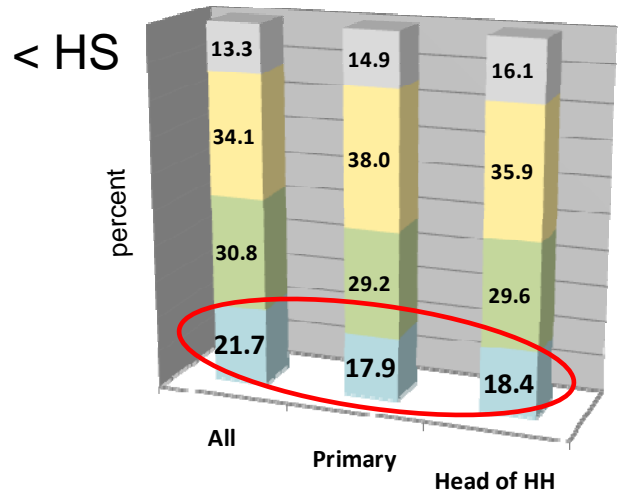
Predicted HH Income vs. KN Actual % by Group



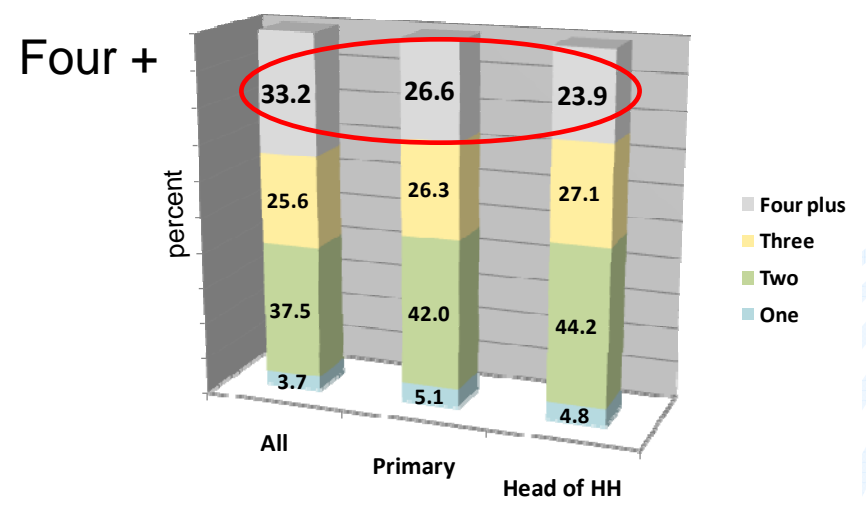
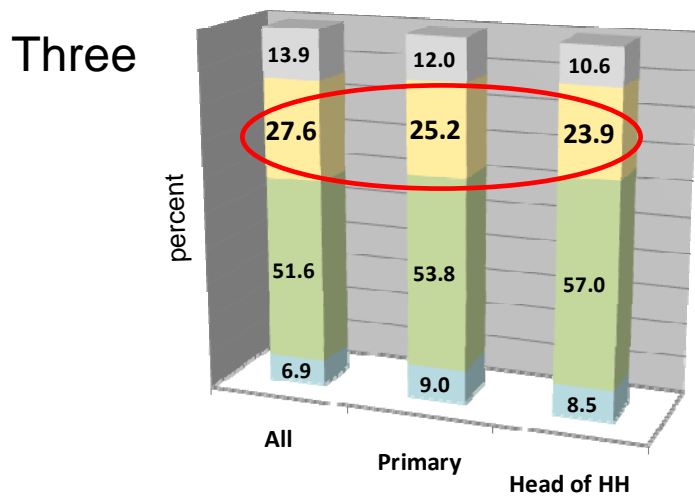
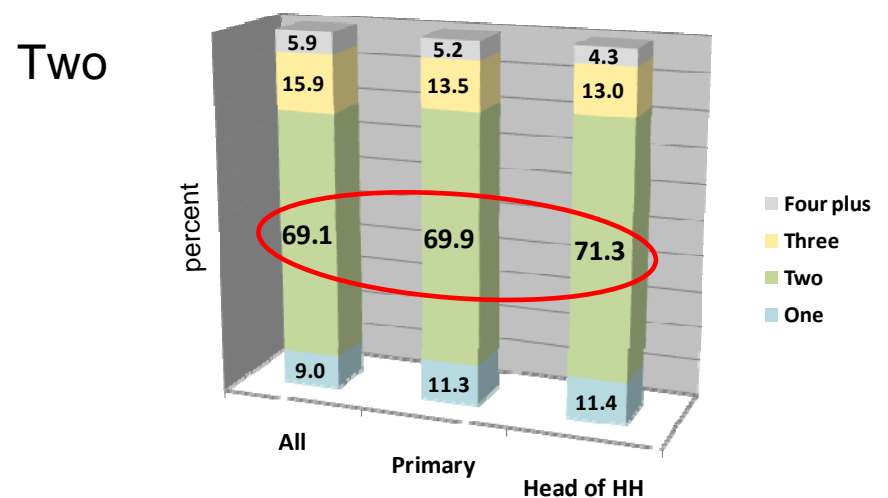
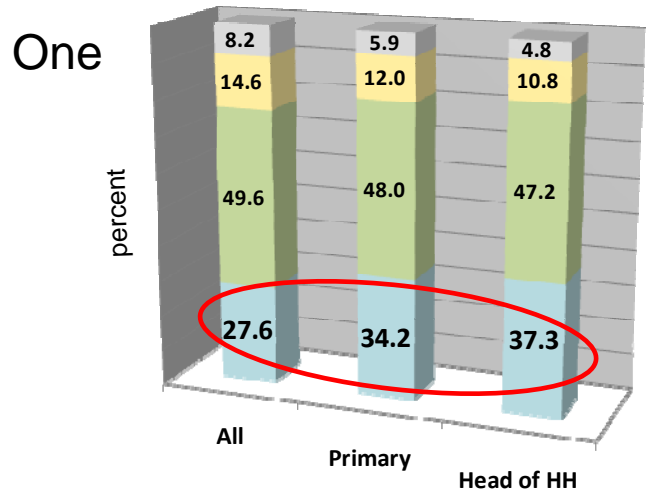
Predicted Children in HH vs. KN actual % by Group



Predicted HOH Education vs. KN Actual % by Group

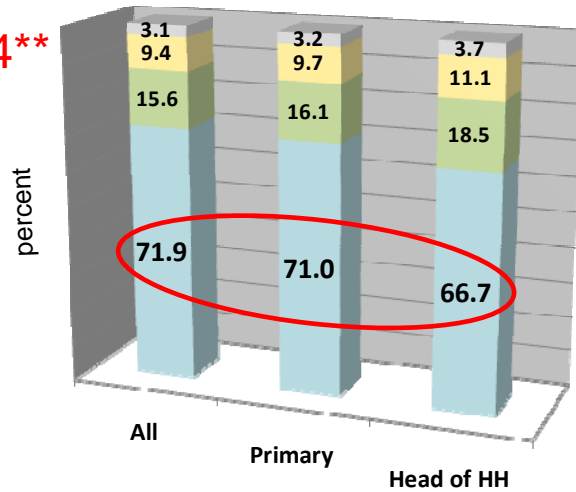


Predicted Number of Adults in HH vs. KN Actual % by Group

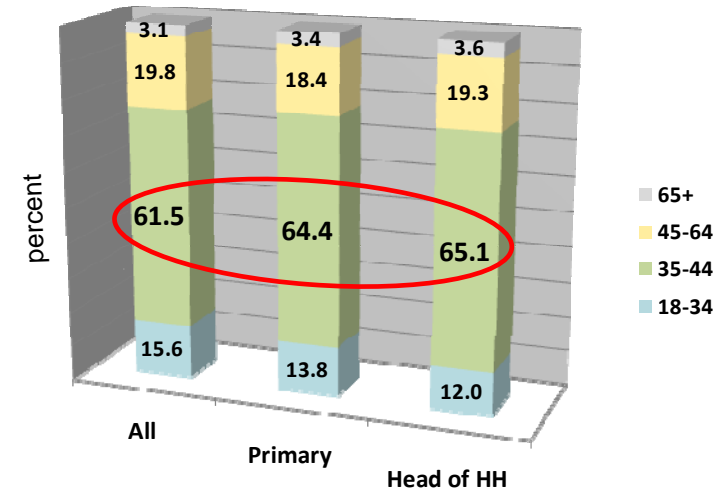


Predicted Age of HOH vs. KN Actual % by Group *

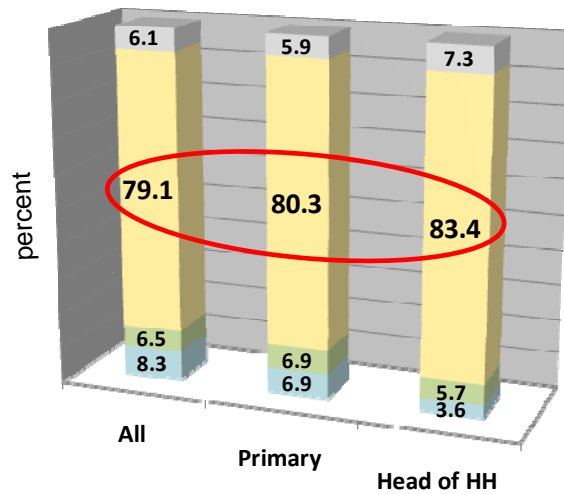
18-34**



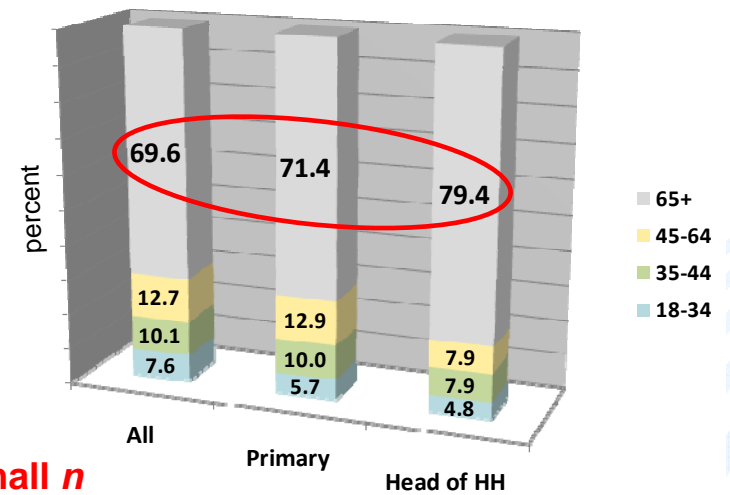
35-44



45-64



65 +



* Pilot data, small n

** Most over age 24

Conclusions

- Ancillary information attached to ABS samples have value for examining response/non-response in surveys and for improving the efficiency of complex sample stratification designs
 - Analysis using additional variables holds promise for future non-response measurement and sample stratification
- Ancillary information appears to correlate best with head-of-household information
- Ancillary information may be useful for mail strategies targeting homeowners, race/ethnic groups, household income extremes, and perhaps with age groups and other groups (pending further research).
 - More research is still needed on the targeting effectiveness for these groups and needs to be extended to other variables (e.g., religion ID).

On the Quality of Ancillary Data Available for Address-Based Sampling

Charles DiSogra

cdisogra@knowledgenetworks.com

(650) 289-2985

Mike Dennis

mdennis@knowledgenetworks.com

(650) 289-2160

Mansour Fahimi

mfahimi@m-s-g.com

(215) 653-7100