UNITED STATES DEPARTMENT OF COMMERCE
Bureau of the Census
Washington, DC 20233-0001

2013 AMERICAN COMMUNITY SURVEY RESEARCH AND EVALUATION REPORT MEMORANDUM SERIES #ACS13-RER-13-R1

DSSD 2013 AMERICAN COMMUNITY SURVEY RESEARCH MEMORANDUM SERIES #ACS13-R-03-R1

MEMORANDUM FOR ACS Research and Evaluation Advisory Group

From: Patrick J. Cantwell ***signed 12/20/13***
Chief, Decennial Statistical Studies Division

Prepared by: Don Keathley
American Community Survey Sample Design Branch
Decennial Statistical Studies Division

Subject: Revision: American Community Survey: Sample Representivity for American Indian Areas

Attached is a revision of the American Community Survey Research and Evaluation report "American Community Survey: Sample Representivity for American Indian Areas", which was originally released in April 2013. This report gives an indication of how representative the interviewed housing units in the American Community Survey (ACS) are of the ACS nonrespondents and, therefore, the sampling frame in American Indian areas. American Indian areas include regions such as Hawaiian homelands and reservations. Representivity is measured using an R-indicator statistic.

Changes from April's release of this document include:

1. A complete revision of Section II.A. – April's release contained factual errors in this section
2. The addition of Section II.D. – it contains a brief of description of the methodology we used for computing standard errors
3. An additional limitation in Section III
4. Revised -2 Log L and Adj R^2 metrics in Table 3

United States Census Bureau

5. The addition of footnotes 8 and 9
6. The inclusion of standard errors in the tables in the attachments, along with some minor revisions in the results section that are associated with their inclusion in this report
7. Some editorial comments throughout the document

None of these changes altered the general results and analysis from April's release.

If you have any questions about this report, please contact Don Keathley (301-763-2225) or Steven Hefter (301-763-4082).

Attachment

cc: ACS Research and Evaluation Workgroup

D. Griffin (ACSO)	M. Asiala	P. Davis	J. Tancreto
M. Ikeda (CSRM)	J. Chesnut	D. Sommers	R. Ramirez (POP)
K. Albright (DSSD)	K. Cyffka	T. Tersine	

DECEMBER 13 , 2013

American Community Survey: Sample Representivity for American Indian Areas

Don Keathley and Steven Hefter
Decennial Statistical Studies Division

United States Census Bureau

Intentionally blank

I. Introduction

The American Community Survey (ACS) has experienced a high response rate since full implementation began in 2005. Overall weighted response rates between 2005 and 2011 range from 97.3 percent in 2005 to 98.0 percent in 2009 (U.S. Census Bureau (2012)). These rates take all three modes of data collection into account (mail, telephone, and personal visit). Vacant housing unit addresses are included in these rates as they are interviews in the ACS. See U.S. Census Bureau (2009) for details.

Although these response rates are high, two to three percent of cases still did not respond. In this evaluation we want to determine whether the nonrespondents are categorically different in any way from the respondents, i.e., are the respondents representative of the nonrespondents and, consequently, of their entire sample? Then, since we assume that each yearly ACS sample is representative of the frame from which it was sampled, we can simultaneously answer the question of whether the respondents are representative of their corresponding frame as well.

The primary statistic we use in measuring representivity is the R-indicator. It is a measure of the spread of response propensities (probabilities of a sample case responding in the survey) across both respondents and nonrespondents. We also look at sample completeness ratios for comparison purposes, which are measures of the combined levels of nonresponse and under- or overcoverage.

Our analysis in this evaluation focuses on American Indian areas only. These areas include regions such as reservations and tribal statistical areas. We estimate sample representivity at the national level as a whole and by various subgroups, e.g., race categories.

II. Metrics

A. R-indicators

Recent years have seen the development of R-indicators. These statistics serve as "indicators" of how well or poorly the respondents of a given survey represent the nonrespondents and, consequently, the population for which the sample represents (we assume that each ACS sample is representative of the sampling frame which, in turn, is representative of the target population). The paper by Skinner, et al. (2009), describes the R-indicators; the paper by Shlomo, et al. (2009) provides a discussion of the statistical properties of the R-indicators; the paper by Schouten, et al. (2009) shows how to apply R-indicators.

Skinner, et al. (2009) and Shlomo, et al. (2009) describe two R-indicators: $R(\rho)$ and q^2, where ρ is a vector of response propensities. We focus on $R(\rho)$ in this paper, due in part to the comment in Schouten, et al. (2009), that "… both indicators lead to similar conclusions about the representativeness of response, although they stem from different objectives," and partly because $R(\rho)$ seems to be the statistic of choice in the literature, e.g., in Schouten.

The R-indicator for the population is defined as

$$R(\boldsymbol{\rho}) = 1 - 2\,S(\boldsymbol{\rho}) \qquad (1)$$

where $\boldsymbol{\rho}$ = vector of response propensities for all units in the population
$S(\boldsymbol{\rho})$ = standard deviation of $\boldsymbol{\rho}$

$$= \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(\rho_i - \bar{\rho})^2} \qquad (2)$$

where N = population size
i = population unit i
ρ_i = response propensity for sample unit i
$\bar{\rho}$ = average response propensity across all sample units
$= \frac{1}{N}\sum_{i=1}^{N}\rho_i$

$S(\boldsymbol{\rho})$ is in the closed interval [0, 0.5]. This means $R(\boldsymbol{\rho})$ is in the closed interval of [0, 1]. $R(\boldsymbol{\rho}) = 1$ when $S(\boldsymbol{\rho}) = 0$, indicating all units in the population have the same propensity to respond. $R(\boldsymbol{\rho}) = 0$ when $S(\boldsymbol{\rho}) = 0.5$, indicating the maximum variation in response propensities.

Equations (1) and (2) are functions of every unit's true propensity to respond – these propensities are usually unknown in practice. When estimating R-indicators in equation (1) from a sample, the response propensities must usually be estimated as well. Equations (3) and (4) define the sample-based R-indicator and standard deviation.

$$\hat{R}(\hat{\boldsymbol{\rho}}) = 1 - 2\,\hat{S}(\hat{\boldsymbol{\rho}}) \qquad (3)$$

where $\hat{\boldsymbol{\rho}}$ = vector of estimated response propensities for the interviewed and noninterviewed sample units from a survey
$\hat{S}(\hat{\boldsymbol{\rho}})$ = standard deviation of $\hat{\boldsymbol{\rho}}$

$$= \sqrt{\frac{1}{N-1}\sum_{i=1}^{n} d_i\,(\hat{\rho}_i - \hat{\bar{\rho}})^2} \qquad (4)$$

where N = population (frame) size
n = sample size
i = sample unit i
d_i = design weight for sample unit i
$\hat{\rho}_i$ = estimated response propensity for sample unit i
$\hat{\bar{\rho}}$ = average estimated response propensity across all sample units
$= \frac{1}{N}\sum_{i=1}^{n} d_i\hat{\rho}_i$

The design weight d_i we used in our computations was the ACS baseweight (BW), where each sample unit's BW is the inverse of its overall probability of selection for sample. We used $\sum_{i=1}^{n} d_i$ in place of N in equation (4).

The passage in Schouten, et al. (2009) above refers to an $\hat{R}(\hat{\boldsymbol{\rho}})$ that is adjusted for bias due to sampling. The $\hat{R}(\hat{\boldsymbol{\rho}})$ and $\hat{S}(\hat{\boldsymbol{\rho}})$ in equations (3) and (4) are unadjusted for this bias[1]. We used equations (3) and (4) due to the large sample sizes in the ACS. As a result, the $\hat{S}(\hat{\boldsymbol{\rho}})$ values we computed are in the left-open interval (0, 0.5]. This means $\hat{R}(\hat{\boldsymbol{\rho}})$ is in the right-open interval of [0, 1).

We estimated response propensities for ACS sample housing units in American Indian areas for the sample years 2007 through 2011 combined. We made these estimates using logistic regression models. The general form of the model is

$$\hat{\rho}_i = e^{g(\mathbf{x_i})} / \left(1 + e^{g(\mathbf{x_i})}\right) \qquad (5)$$

where $g(\mathbf{x_i})$ is a linear regression function, i.e., $\beta_0 + \beta_{1i}x_{1i} + \ldots + \beta_{ki}x_{ki}$, where k is the number of regressors in the model.

When transformed via a natural logarithm, $g(\mathbf{x_i})$ in equation (5) becomes

$$g(\mathbf{x_i}) = \ln\left[\frac{\hat{\rho}_i}{1-\hat{\rho}_i}\right] \qquad (6)$$

The regressors are variables for which all responding and nonresponding sample units have a value. These variables are referred to as sample-based auxiliary information in, e.g., Skinner, et al. (2009). We assume that this information comes from one or more sources external to the survey in question, such as administrative record data. Regressors were chosen that we found to have a strong correlation with the survey's response propensities. We chose the variables listed in Table 1 as the regressors.

Most of the regressors are unit-level 2010 Census variables from the 2010 Census Hundred-Percent Detailed File (HDF) for housing units, while two come from geographic reference files and one from a Geography Division-supplied file (see Attachment A for file descriptions).

We ran standard weighted stepwise logistic regressions to determine which of the regressors are significant and to help us decide which variables are worth keeping in the model[2]. Our weights were the design weights (d_i) from above. The dependent variable is a binary response indicator (RI), where $RI_i = 1$ if ACS sample unit i responded and 0 if unit i did not respond.

B. Sample Completeness Ratios (SCR)

An adjunct to $\hat{R}(\hat{\boldsymbol{\rho}})$ is the sample completeness ratio (SCR – see Albright and Starsinic (2002)). It is the ratio of the sum of the baseweights (design weights) of the responding sampled units in the survey divided by an independent count or control.

[1] Both Schouten, et al. (2009) and Shlomo, et al. (2009) indicate that biases would be downward, meaning the adjusted $R(\boldsymbol{\rho})$ values would be higher than their unadjusted counterparts.

[2] This includes computations of standard errors for parameters, i.e., we did not use the successive difference replication method that the ACS uses for its estimates.

These weights take personal interview sub-sampling into account. The general equation for an SCR is

$$SCR = (\sum_r d_r)/(N) \quad (7)$$

where r = ACS respondent r
d_r = design weight for ACS respondent r
N = independent total

We compute SCRs at the national level and at the sub-national level for given variables (most of the regressors in Table 1). The general equation we use in this evaluation is

$$SCR_{c,HDF} = (\sum_r d_{c,HDF,r})/(N_{c,HDF}) \quad (8)$$

where c = category/characteristic c (category variable value)
HDF = source of auxiliary data
$SCR_{c, HDF}$ = SCR for category c, where the source for category c classifications is the HDF
$d_{c, HDF, r}$ = design weight for ACS respondent r that matched to the HDF, in category c
$N_{c, HDF}$ = count of cases on the ACS frame that matched to the HDF, in category c

For example, if we computed the national SCRs for each householder's race category, then each ACS respondent's value for race will come from the HDF and the independent controls will be the counts of householders on the HDF for which a match could be made to the ACS frames for each race category – the equation is

$$SCR_{race\ c,\ HDF} = (\sum_r d_{race\ c,HDF,r})/(N_{race\ c,HDF}) \quad (9)$$

SCRs show the proportions of the universe/frame that is represented by the respondents, before any adjustments (e.g., for nonresponse) are made to the respondents' design weights, i.e., it is not an indicator of sample respondent representivity. What they do indicate is the magnitude of nonresponse, under- or overcoverage, or both that are present in the sample. An SCR = 1 is the ideal situation – it means 100 percent coverage and, potentially 100 percent response. Any deviation from one indicates the presence of nonresponse, under- or overcoverage, or both.

C. R-Indicators and SCRs

The best-case scenario is when the R-indicator is just less than one and the SCRs are equal to one. This would show almost perfect sample representativeness combined with 100 percent coverage, response, or both. We continue to assume in the research that the sample is representative of the frame.

Lower-valued R-indicators indicate some degree of a lack of sample representativeness. Should the SCRs be close to or equal to one, however, then this lack of representativeness might not be an issue except for a small segment(s) of the population, e.g., an age group.

R-indicators close to one show good representativeness of the respondents relative to the nonrespondents and the frame. If SCRs are relatively small, however, then the frame (and therefore the sample) might not necessarily be representative of the target population.

The worst-case scenario is when both lower-valued R-indicators and relatively small or large SCRs occur. This result would indicate both a certain lack of sample representativeness combined with the possibility of the frame not being representative of the target population.

D. Standard Error Estimates

The R-indicators and SCRs are both estimates based on samples (the ACS in our case). This means they are both subject to sampling error. The ACS uses the successive difference replication (SDR) method for computing standard errors for its estimates – we do the same for the R-indicators and SCRs. The general SDR equation we use is

$$SE = \sqrt{\frac{4}{80}\sum_{i=1}^{80}\left(X_i - X\right)^2} \qquad (10)$$

where X_i = estimate (R-indicator, SCR) from replicate sample i, $i \in \{1, \ldots, 80\}$
X = estimate (R-indicator, SCR) from the base sample

Each ACS sample unit has a set of eighty replicate factors. We multiplied every sample unit's final baseweight by each of its replicate factors, resulting in eighty replicate samples. For the R-indicators, we ran each replicate sample through the best-fitting models below, resulting in eighty sets of R-indicators (total and by category, for the variables in Table 1). We then applied equation (10) to obtain the standard errors for the R-indicators. For the SCRs, we first computed eighty sets of numerators by summing each replicate sample's adjusted baseweights, across all units and by variable category (from Table 1). Then we computed the SCRs by dividing the replicate sample numerators by the appropriate denominators (the denominators are from the base sample for each replicate SCR). We then applied equation (10) to get standard errors for the SCRs.

See Ash (2011) for details on the SDR method.

III. Limitations

One limitation is that our analysis was restricted to just those ACS interviews and noninterviews with a MAF (Master Address File) Tiger Feature Class Code (MTFCC)

shown in Attachment E (eligible case) that matched to the HDF. Approximately 9.5 percent of the eligible cases did not match to the HDF (24,540 of 258,423 cases)[3] [4]. And, if the average weighted response rates are different between the matching versus nonmatching cases for one or more MTFCC categories, then these differences could have an impact on the SCRs we actually observe had all eligible cases matched to the HDF.

Another limitation is that not all of the matching eligible cases had entries for the variables on the HDF, i.e., many were vacants in the 2010 Census. These records comprised approximately 5.5 percent of the matching eligible cases (12,921 of 233,883 cases).

If all of the nonmatches had matched to the HDF and if all of the matches had been occupied housing units in 2010, then our results might have been different from those observed, for both the R-indicators and the SCRs.

Another limitation is that the matching was done by MAFID only. MAFIDs might not always refer to the exact same address across time. Had the HDF contained address information, like house number and street name, then matching could have been performed using these variables. This would have potentially resulted in more accurate matching between the files.

One more limitation is that it is possible that the HDF values for the matching ACS sample cases might be different than what was reported in the ACS.

IV. Methodology

A. Input Files, Variables (Regressors)

Table 1 shows the variables we used as a basis for our regressors, along with their source files. It also shows the source for the dependent variable (STATUS / ACSINT). See Attachment A for descriptions of all of the files mentioned in this section.

We merged various files, including those shown in Table 1, to create the input file for the logistic regression modeling and R-indicator computations. This file contains all of the variables shown in Table 1. Attachment B provides a summary of the process we used to create the final input file.

The codes for each variable that we used are shown in the table in Attachment C. The last column in the table shows the code/category we used as the reference group for the regressor[5].

[3] The 258,423 total excludes matches that were not housing units in the 2010 Census – there were 17 such cases.

[4] One possible cause for the nonmatches is that some of the ACS sample housing unit addresses from 2007 to 2011 may not have not have existed during the 2010 Census.

[5] Reference groups are the levels of the variables in a model against which the parameter estimates for the remaining levels are compared.

Table 1. Variables, Source Files

	Description	Source File
	Edited Building Structure Type	2010 Unit HDF *
CLUSTERNUM / SEG_GRP	Segmentation Group Code	File from Geography Division
	Edited Age of Householder	2010 Unit HDF
HHSPAN	Hispanic or Latino Householder	2010 Unit HDF
	Race of householder	2010 Unit HDF
HHT	Household Family Type	2010 Unit HDF
	Legal/Statistical Area Definition Code	2007-2011 GRFC #, GRFN &
MTFCC	MAF Tiger Feature Class Code	2007-2011 GRFN
	ACS Interview Outcome Code	2007-2011 Select Files
TENSHORT	Tenure	2010 Unit HDF

* The 2010 Unit HDF is the housing-unit level data file from the 2010 Census, where the data are edited
# Geographic Reference Files, with coded geography
& Geographic Reference Files, with named geography

We copied the variable CLUSTERNUM to SEG_GRP, with a recode: CLUSTERNUM = blank became SEG_GRP = 0 This was done for programming purposes, where a blank was not an acceptable value. We recoded STATUS to ACSINT so that ACS interviews and non-interviews had codes of 1 and 0, respectively.

B. Logistic Regression Models

We ran two sets of logistic regression models, one which included the variable LSADC and one including MTFCC. These models are shown in Table 1. We ran these models using only those ACS interviews and noninterviews with an MTFCC that was equal to one of the codes in Attachment E (eligible cases). Each eligible case had an LSADC equal to one of the codes shown in Attachment D.

All of the models except 5L were exploratory models, where we compared the models to each other, primarily with respect to model fit. Model 5L took the best fitting of the six models (1L – see Table 3) and collapsed the American Indian areas with parameters from model 1L that were not significant into one parameter.

All were regular stepwise regression models, and all were weighted using the sampled units' design weights (baseweights). The significance level cutoff for inclusion in the model was 0.01. We ran the models using housing unit records for which we had entries for the variables only, i.e., for which the housing unit was occupied in the 2010 Census – non-vacants[6].

[6] The ACS classifies all vacant units as interviews. If we had had information on age, sex, etc. for the householders of these units, at least some of the R-indicator values we observed would have moved closer to 1.

Table 2. Models

Model	Description
1L	All main effects only from Table 1, except MTFCC, using cases with acceptable LSADC values only
1M	All main effects only from Table 1, except LSADC, using cases with acceptable MTFCC values only
2L	Same as 1L, except with two-way interactions
2M	Same as 2M, except with two-way interactions
3	Same as 1L/1M – all main effects except MTFCC/LSADC
4	Same as 2L/2M – all mail effects except MTFCC/LSADC and all interactions except those involving MTFCC/LSADC
5L	Same as 1L, except collapsed only those LSADC categories that had non-significant parameters

C. R-Indicators

Once we completed the logistic regression runs, we used equations (3) and (4) to calculate the values of $\hat{R}(\hat{\rho})$ from each logistic regression run.

D. Sample Completeness Ratios

We computed SCRs for totals and main effects. Since the logistic regression models and R-indicators are based on 2010 Census occupied housing units only, we compute SCRs for occupied housing units only as well (exception: we compute a national SCR for vacants). The numerators are weighted summations from records in the final input file mentioned in Section IV.A. The denominators are counts of matching records between the HDF and the five yearly ACS sample frames (the edited MAF extracts). Matching was on the nine-digit MAFID (the 2007 edited MAF extracts include the twelve-digit MAFIDs only, so we added their nine-digit MAFIDs from the 2010 edited MAF extracts. We matched the 2007 and 2010 extracts on the twelve-digit MAFID).

E. Model-Fit Metrics

Table 3 shows summaries of the results from each model. The goodness-of-fit metrics are indicators of how well each model fits in comparison to the other models. -2 Log L is -2 times the log-likelihood of the model, where lower values indicate better fits[7].

Adjusted (Adj) R^2 (Nagelkerke (1991)) is the ratio of a generalization of the coefficient of determination (CD) divided by its maximum possible value:

[7] We looked at the Akaike Information Criterion (AIC) as well – we omit this statistic because the values we observed for all models was approximately the same as that for -2 Log L.

$$\text{Adj } R^2 = R^2 \,/\, \text{Max } R^2 \qquad (11)$$

where R^2 = a generalization of the CD (Cox and Snell (1989))

$$= 1 - \left(\frac{L(0)}{L(\hat{\beta})}\right)^{2/n} \qquad (12)$$

Max R^2 = maximum R^2 value

$$= 1 - (L(0))^{2/n} \qquad (13)$$

$L(0)$ = log-likelihood of the intercept-only model
$L(\hat{\beta})$ = log-likelihood of the specified model
n = weighted sample size

The reason for using Adj R^2 is that its maximum value is one, whereas it is less than one for R^2 (both statistics can take on minimum values of zero). Higher values of Adj R^2 indicate a better model fit.

The receiving operating characteristic (ROC) curve is a plot of proportions of true positive predictions (sensitivity) on the y-axis versus proportions of false positive predictions (1 – specificity) on the x-axis, at various sensitivity levels. The sensitivity levels range from zero to one, inclusive. Each level indicates the proportion of true positives that are classified as positives by the model, given a probability cut-off point; in our case, positives are interviews and negatives are noninterviews. For example, if the cut-off point is 80 percent, then any case with a predicted probability greater than or equal to 80 percent is classified as a positive (interview, in our case); those with a predicted probability less than 80 percent are classified as a negative (noninterview). The sensitivity level is then the proportion of true positives (interviews) that are classified as positives (interviews) given the 80 percent cut-off point. The false positive (interview) rate associated with a given sensitivity level indicates the proportion of true negatives (noninterviews) that are classified as positives (interviews) given the cut-off point. Thus, the ROC curve for each of our models is a plot of proportions of true interview classifications versus false interview classifications.

The area under the ROC curve indicates how well a model differentiates between true positives (interviews) and true negatives (noninterviews). An area of one shows perfect predictions, or discrimination, in the model – all of the cases that are predicted to be positive at any given sensitivity level are true positives. An area of 0.5 indicates zero discrimination – half of the cases that are predicted to be positive at any sensitivity level are true positives and half are true negatives. As areas increase from 0.5 to 1, the ability of the model to discriminate between true positives and negatives increases. Areas less than 0.5 indicate a negative discrimination, where more than half of the cases predicted to be positive are actually true negatives. See Kleinbaum and Klein (2010) for more information on ROC curves.

V. Results

The results in Table 3 show that, of the models that are not questionable (1L, 1M, 3, 4, and 5L), models 1L and 5L fit the best: they have the largest Adj R^2 and ROC curve areas. They have the smallest -2 Log L values as well, but the percent differences between their values and those for models 1M, 3, and 4 are minimal. Of concern, however, are the relatively small ROC curve areas – the 0.693 values indicate fits that could be questionable, since they are closer to 0.5 than 1.0.

Table 3. Summary of Logistic Regression Runs – American Indian Areas

			Goodness-of-Fit Metrics		
Model	**Steps**	**Variables in Model**	**-2 Log L**	**Adj R^2**	**Area under ROC Curve**
1L	8	All	419,201	0.087	0.693
1M	8	All	421,627	0.075	0.677
2L	36	All, but the model fit was questionable after step 3	435,491	0.005	0.725
2M	36	All, but the model fit was questionable after step 4	433,109	0.017	0.711
3	7	All	426,223	0.052	0.663
4	28	All, but model fit was questionable after step 15	425,535	0.056	0.667
5L	8	All	419,281	0.087	0.693

Sources: 2010 Census data, American Community Survey interviews and non-interviews from 2007-2011

We still computed $\widehat{R}(\hat{\rho})$ values for each model, which are shown in Table 4.

Table 4. $\widehat{R}(\hat{\rho})$ Values – American Indian Areas

	$\widehat{R}(\hat{\rho})$, (S.E.)^	Model	$\widehat{R}(\hat{\rho})$, (S.E.)^
	0.965 (0.0035)	3	0.974 (0.0014)
1M	0.968 (0.0016)	4	0.935 (0.0250)
	0.854 (DNC*)	5L	0.965 (0.0016)
2M	0.867 (DNC)		

Sources: 2010 Census data, American Community Survey interviews and non-interviews from 2007-2011

^ S.E. = standard error

* DNC = Did not compute

Except for models 2L and 2M, all of the $\widehat{R}(\hat{\rho})$ values are between 0.935 and 0.974. Models 1L and 5L, the models with the best fit, have $\widehat{R}(\hat{\rho})$ values of 0.965**.**

The tables in Attachment F show $\widehat{R}(\hat{\rho})$ values for model 5L for each category for each main effect. We chose model 5L over 1L due to coefficient of the collapsed LSADC being

significant in 5L, but any differences between the two models were minimal. All but one of the $\widehat{R}(\hat{\rho})$ values in model 5L were statistically significantly greater than 0.940 (exception: BLD = O, other type of building structure (Table F-1) – $\widehat{R}(\hat{\rho})$ = 0.942, with a standard error = 0.024).

SCRs for totals are shown in Table 5; the tables in Attachment F show SCRs for the main effects. SCRs are shown only for cases with MTFCC values shown in Attachment E. All SCRs in Attachment F omit vacant units.

Table 5. Sample Completeness Ratios for Totals – American Indian Areas

Regressor	SCR (S.E.)^
Total	0.836 (0.0018)
Total, minus vacants	0.907 (0.0020)

Sources: 2010 Census data, American Community Survey interviews and non-interviews from 2007-2011

^ S.E. = standard error

Table 5 and those in Attachment F show SCRs greater than 0.85 for most categories, and above 0.800 for all but four categories (BLD = O – other type of building structure (SCR = 0.627, Table F-1), MTFCC = G2200: ANRC (SCR = 0.711, Table F-8), and SEG_GRP ∈ {4, 6} – (economically disadvantaged II, renter skewed (SCR = 0.831[8]) and ethnic enclave II, renter skewed tracts, respectively (SCR = 0.746, Table F-6))[9]. Two categories had SCRs that were statistically significantly greater than one (SEG_GRP = 0: no segmentation group code (SCR = 1.024, Table F-6) and MTFCC = G2130: ANVSA (SCR = 1.518, Table F-8)).

VI. Conclusions

Given the proximity of the majority of $\widehat{R}(\hat{\rho})$ values to one, we would infer that the sample respondents in American Indian areas as a whole and by main effect category are fairly representative of their corresponding sample nonrespondents. Since we assume that the sample itself is representative of the frame, we would also infer that the sample respondents are representative of the frame as well.

The overall SCR for total, minus vacants in Table 5 (0.907) combined with the $\widehat{R}(\hat{\rho})$ value for models 1L and 5L (0.965) indicates that the sample respondents in American Indian areas as a whole are fairly representative of the target population in these areas. This goes for many main effect categories in Attachment F as well, e.g., American Indians/Alaska Natives in Table F-4 ($\widehat{R}(\hat{\rho})$ = 0.953, SCR = 0.931). Some of the SCRs are relatively small, however (e.g., SCR = 0.711 for Alaska Native Regional Corporations (MTFCC = G2200)), suggesting that the respondents for some categories may not be entirely representative of their target populations.

[8] This SCR was not statistically significantly different from 0.8.

[9] The SCRs for SEG_GRP = 6 and MTFCC = G2200 were not statistically significantly different from each other.

VII. Future Research

Future research could include the use of ACS data, auxiliary information from other sources, or both. Examples of other auxiliary information sources are the Census Bureau's planning database and Internal Revenue Service records. We would potentially have a higher proportion of ACS sample cases with complete auxiliary information from alternate sources than we did for this analysis.

We could conduct this research for subsets of the ACS samples, e.g., ACS data collection mode and by ACS sampling stratum. It is possible that representivity could fluctuate between modes or strata, or both.

Other research could explore the use of the bias-adjusted $\widehat{R}(\hat{\rho})$ and the q^2 R-indicators mentioned in Schouten, et al. (2009), as comparisons to the results in this report. We could also compare the standard errors we computed with those from a Taylor linearization method from the literature, for comparison purposes.

Some additional research could include looking into why some groups have outlier R-indicator and SCR values, e.g., for BLD = O values. Matching the ACS and Census records on address information, while more involved, would allow us to compare the results of this matching with the matching we did for this evaluation (by MAFID). We could compute R-indicators across time, e.g., on a yearly basis, as a monitoring device.

VIII. References

Albright, K. and Starsinic, M. (2002), "Coverage and Completeness in the Census 2000 Supplementary Survey," *Proceedings of the Section on Survey Research Methods, American Statistical Association,* 3,345-3,349

Ash, S. (2011), "Using Successive Difference Replication for Estimating Variances," *Proceedings of the Section on Survey Research Methods, American Statistical Association,* 3,534-3,548

Boone, T. (2008), "Segmenting the Population for the 2010 Census Integrated Communications Campaign," C2PO 2010 Census Integrated Communications Research Memoranda Series Number 1, available at http://www.census.gov/2010census/partners/pdf/C2POMemoNo_1_10-24-08.pdf, last accessed in December 2013

Cox, D.R. and Snell, E.J. (1989), *The Analysis of Binary Data (2nd Edition),* London: Chapman and Hall

Keathley, D. and Hefter, S. (2013), "Revision: American Community Survey: R Indicators for the Nation and Puerto Rico," 2013 American Community Survey Research and Evaluation Report, Series Number ACS13-RER-12-R1

Kleinbaum, D. and Klein, M. (2010), *Logistic Regression (3rd Edition)*, New York: Springer

Nagelkerke, N.J.D. (1991) "A Note on a General Definition of the Coefficient of Determination," *Biometrika*, 78, 691-692

Schouten, B., Morren, M., Bethlehem, J., Shlomo, N., Skinner, C. (2009), "How to use R-indicators?", Work Package 4, Deliverable 3, RISQ Project, 7th Framework Programme (FP7) of the European Union, available at http://www.risq-project.eu/papers/RISQ-Deliverable-3.pdf, last accessed in December 2013

Shlomo, N., Skinner, C., Schouten, B., Bethlehem, J., Zhang, L. (2009), "Statistical Properties of R-indicators," Work Package 3, Deliverable 2.1, RISQ Project, 7th Framework Programme (FP7) of the European Union, available at http://www.risq-project.eu/papers/RISQ-Deliverable-2-1-V2.pdf, last accessed in December 2013

Skinner, C., Shlomo, N., Schouten, B., Zhang, L., Bethlehem, J. (2009), "Measuring Survey Quality through Representativeness Indicators using Sample and Population Based Information", *Paper presented at the NTTS Conference, 18-20 February 2009, Brussels, Belgium*, available at http://www.risq-project.eu/papers/skinner-shlomo-schouten-zhang-bethlehem-2009-a.pdf, last accessed in December 2013

U.S. Census Bureau (2009), "(ACS) Design and Methodology," available at http://www.census.gov/acs/www/Downloads/survey_methodology/acs_design_methodology.pdf, last accessed in December 2013

U.S. Census Bureau (2012), "Response Rates – Data": http://www.census.gov/acs/www/methodology/response_rates_data/, initially accessed in August 2012

Table A. Input, Output Files

File	Description
ACS Sample File	File used as input to the logistic regression models and for the numerators in the SCR equations
Edited MAF Extracts (EDMAF)	Edited MAF extracts that have been through ACS edits and code assignments; used as inputs for ACS sampling.
EDMAF-HDF Match File	Sample-year files containing matching records between the edited MAF extracts for the given year and the HDF. Used to compute the denominators in the SCR equations.
Geographic Reference File – Codes	Files that contain block-level geographic codes, e.g., codes for legal/statistical area descriptions.
Geographic Reference File – Names	Files that contain names for the geographic entities in the codes files, except block and "filler" codes.
Hundred-Percent Detail File (HDF)	A file containing edited characteristics and records for all households in the 2010 Census. The data have also been through a disclosure avoidance and tabulation geography application.
Sample Delivery File	Final sample files sent to the American Community Survey Office, as inputs to their sample control system. They are subsets of the second-stage sample files, containing valid records only.
Segmentation Group File	A tract-level file containing segmentation group (CLUSTERNUM) codes for each applicable tract.
Select File	Files that contain the final interview status code for ACS sample housing unit addresses.
Second-Stage Sample File	Output files from the housing unit address sample selection process. They include invalid records.

Summary of the Input File Creation Process

Note: All of the files in this summary are shown in the table in Attachment A.

We started by creating five files that contained one record per ACS sample housing unit address (ACSSAMP). Each file contained the sampled addresses for one of the five sample years in which we were interested, i.e., 2007 through 2011. Each file was a concatenation of the corresponding year's sample delivery files. There are eight sample delivery files per year, four for the United States and four for Puerto Rico.

We added interview status for each sampled address by matching each ACSSAMP file to its corresponding year's select file, on CMID (nine-digit continuous measurement id). Then we matched the ACSSAMP files to their corresponding second-stage sample files, also on CMID, to pick up each sampled address' baseweight, second-stage sampling stratum, CAPI sub-sampling stratum, reduction measure-of-size, and some geography variables.

We then merged the ACSSAMP files to each corresponding year's geographic reference files–codes (GRFC), to pick up the Alaska Native Regional Corporation (ANRC) code (ANRCCE in 2007 and 2008, ANRCFP in 2009, 2010, and 2011) for each sampled address in Alaska that was in an ANRC. We did this matching only for those areas where the American Indian Area code (AINDN; is referred to as AIANHH in the GRFC documentation) was blank (ANRC-only areas), as those with filled AINDN codes were also in Alaska Native Village Statistical Areas, and we wanted to code them as such. Matching was at the block level.

We picked up legal/statistical designation codes (LSADC) for ANRC-only areas by matching the ACSSAMPs to their corresponding year's geographic reference file-names (GRFN). Matching was done on a state -by- ANRCCE/ANRCFP level. We picked up LSADCs for the remaining sampled addresses from the GRFNs as well. Matching for these cases was on state -by- American Indian area -by- tribal subdivision level.

The foregoing process of matching to the GRFC and GRFN files was necessary due to the ANRC values for the LSADC not having been present on any ACS sample files.

The MAF (Master Address File) Tiger Feature Class Code (MTFCC) variable that we needed was already present on the GRFN files for 2009, 2010, and 2011. They did not exist on the 2007 and 2008 GRFNs, however – we used the variables record type (RT) and American Indian Area code (AINDN) from the GRFNs to create MTFCCs for sampled records in these two years.

Summary of the Input File Creation Process (continued)

We obtained the variable CLUSTERNUM by matching the ACSSAMP files to a segmentation cluster file that was created by geography division. This file contained one record per tract. Matching between the two files was at the tract level. Not all tracts are represented on the cluster file, so some records in ACSSAMP did not have a segmentation group code.

Finally, we merged each ACSSAMP to the 2010 Census unit-level hundred-percent detail file (HDF in Table 1). The matching was done on the nine-digit MAFID code. Since MAFIDs in 2007 were the old twelve-digit versions, we needed to match the 2007 ACSSAMP to the 2010 supplemental edited master address files to pick up the 2007 sample's nine-digit MAFIDs prior to matching to the HDF.

The final ACSSAMP files contain only those sample records that matched to the HDF This is because non-matching records from the ACSSSAMPs would not have any data for the majority of the independent variables in the logistic regression models

The actual input file to the logistic regression modeling and R-indicator computations is a concatenation of the final individual year ACSSAMP files.

Table C. Variable Values for the Regressors

Variable	Regressor	Values	Reference Group
BLD	BLD	S = one-family house M = multi-family house T = trailer/mobile home O = other (boat/RV/van, etc)	S
CLUSTERNUM	SEG_GRP	See Attachment G	0
HHLDRAGE	AGE	1 = 0 to 24 2 = 25 to 34 3 = 35 to 44 4 = 45 to 54 5 = 55 to 64 6 = 65 to 74 7 = 75+	2
HHSPAN	HHSPAN	1 = not Hispanic or latino 2 = Hispanic or latino	1
HHRACE	RACE	1 = White alone 2 = Black alone 3 = Amerind/Alaskan Native alone 4 = Asian alone 5 = Native Hawaiian/pacific islander alone 6 = Some other race alone 7 = Multi-race	1
HHT	HHT	1 = Husband/wife family household 2 = Other family household: male householder 3 = Other family household: female householder 4 = Nonfamily household: male householder, living alone 5 = Nonfamily household: male householder, not living alone 6 = Nonfamily household: female householder, living alone 7 = Nonfamily household: female householder, not living alone	1
LSADC	LSADC	See Attachment D	00
MTFCC	MTFCC	See Attachment E	G2100
STATUS	ACSINT	1 = Interview (ACSINT = 1) 4 = Non-Interview (ACSINT = 0) All other codes were out-of-scope for this evaluation	-
TENSHORT	TENSHORT	1 = Owner-occupied unit 2 = Renter-occupied unit	1

Table D. Legal/Statistical Area Description Codes (LSADC) for American Indian Areas

LSADC	LSADC Description
00	Blank (2009-2011 only)
07	District (tribal sub-division; 2008-2011 only)
28	District (tribal sub-division)
77	Alaska Native Regional Corporation (ANRC)
78	Hawaiian Home Land
79	Alaska Native Village Statistical Area (ANVSA)
80	Tribal Designated Statistical Area (TDSA)
81	Colony
82	Community (tribal sub-division)
83	Joint-use area
84	Pueblo
85	Rancheria
86	Reservation
87	Reserve
88	Oklahoma Tribal Statistical Area (OTSA)
89	Trust Land
90	Joint-use OTSA
91	Ranch
92	State Designated Tribal Statistical Area (SDTSA)
93	Indian Village
94	Village
95	Indian Community
96	Indian Reservation
97	Indian Rancheria
98	Indian Colony

Table D. Legal/Statistical Area Description Codes (LSADC) for American Indian Areas (continued)

LSADC	LSADC Description
99	Pueblo de (prefix of specific entity name)
9A	Blank (tribal sub-division)
9B	Blank (tribal sub-division)
9C	Pueblo of (prefix of specific entity name)
9D	Settlement
9E	Rancheria Reservation
9F	Ranches
IB	Tribal Block Group
IT	Tribal Census Tract
OT	Off-reservation Trust Land
T1	Area (tribal sub-division)
T2	Chapter (tribal sub-division)
T3	Segment (tribal sub-division)
T4	Blank (tribal sub-division)
T5	Blank (tribal sub-division)
T6	Blank (tribal sub-division)
TA	Administrative Area (tribal sub-division)
TB	Addition (tribal sub-division)
TC	County District (tribal sub-division)
TD	Sector (tribal sub-division)

Table E. MAF (Master Address File) Tiger Feature Class Code (MTFCC) for American Indians

MTFCC	MTFCC Description
G2100	Legal American Indian Area
G2120	Hawaiian Homeland
G2130	Alaska Native Village Statistical Area (ANVSA)
G2140	Oklahoma Tribal Statistical Area (OTSA)
G2150	State Designated Tribal Statistical Area (SDTSA)
G2160	Tribal Designated Statistical Area (TDSA)
G2170	Joint-use Area
G2200	Alaska Native Regional Corporation (ANRC)
G2300	Tribal Subdivision

Table F-1. $\hat{R}(\hat{\rho})$ and SCR Values for Edited Building Structure Type (BLD)

Edited Building Structure Type	$\hat{R}(\hat{\rho})$, (S.E.)^	SCR, (S.E.)^
One-family house	0.969 (0.0015)	0.924 (0.0024)
Multi-family house	0.959 (0.0033)	0.858 (0.0073)
Trailer/mobile home	0.964 (0.0021)	0.840 (0.0074)
Other (boat/RV/van, etc.)	0.942 (0.0237)	0.627 (0.0534)

Sources: 2010 Census data, American Community Survey interviews and non-interviews from 2007 through 2011
^ S.E. = standard error

Table F-2. $\hat{R}(\hat{\rho})$ and SCR Values for Edited Age of Householder (AGE)

Edited Age of Householder	$\hat{R}(\hat{\rho})$, (S.E.)^	SCR, (S.E.)^
0 to 24	0.966 (0.0038)	0.828 (0.0130)
25 to 34	0.961 (0.0027)	0.891 (0.0071)
35 to 44	0.963 (0.0024)	0.902 (0.0063)
45 to 54	0.965 (0.0021)	0.914 (0.0058)
55 to 64	0.970 (0.0020)	0.915 (0.0060)
65 to 74	0.977 (0.0018)	0.924 (0.0060)
75+	0.982 (0.0015)	0.927 (0.0077)

Sources: 2010 Census data, American Community Survey interviews and non-interviews from 2007 through 2011
^ S.E. = standard error

Table F-3. $\hat{R}(\hat{\rho})$ and SCR Values for HHSPAN (Hispanic or Latino Householder)

Hispanic or Latino Householder	$\hat{R}(\hat{\rho})$, (S.E.)^	SCR, (S.E.)^
Not Hispanic or Latino	0.966 (0.0016)	0.907 (0.0020)
Hispanic or Latino	0.964 (0.0040)	0.919 (0.0120)

Sources: 2010 Census data, American Community Survey interviews and non-interviews from 2007 through 2011
^ S.E. = standard error

Table F-4. $\hat{R}(\hat{\rho})$ and SCR Values for Race of Householder (RACE)

Race of Householder	$\hat{R}(\hat{\rho})$, (S.E.)^	SCR, (S.E.)^
White alone	0.973 (0.0017)	0.907 (0.0021)
Black alone	0.970 (0.0027)	0.876 (0.0069)
American Indian /Alaska Native alone	0.953 (0.0029)	0.931 (0.0068)
Asian alone	0.962 (0.0076)	0.893 (0.0250)
Native Hawaiian / Pacific Islander alone	0.974 (0.0074)	0.945 (0.0498)
Some other race alone	0.969 (0.0036)	0.931 (0.0188)
Multi-race	0.961 (0.0038)	0.916 (0.0149)

Sources: 2010 Census data, American Community Survey interviews and non-interviews from 2007 through 2011
^ S.E. = standard error

Table F-5. $\hat{R}(\hat{\rho})$ and SCR Values for Household Family Type (HHT)

Household Family Type	$\hat{R}(\hat{\rho})$, (S.E.)^	SCR, (S.E.)^
Husband/wife family household	0.976 (0.0016)	0.928 (0.0030)
Other family/household: male householder	0.964 (0.0029)	0.913 (0.0114)
Other family household: female householder	0.968 (0.0021)	0.898 (0.0068)
Nonfamily household: male householder, living alone	0.953 (0.0032)	0.856 (0.0072)
Nonfamily household: male householder, not living alone	0.971 (0.0037)	0.877 (0.0153)
Nonfamily household: female householder, living alone	0.965 (0.0026)	0.899 (0.0069)
Nonfamily household: female householder, not living alone	0.964 (0.0055)	0.891 (0.0164)

Sources: 2010 Census data, American Community Survey interviews and non-interviews from 2007 through 2011
^ S.E. = standard error

Table F-6. $\hat{R}(\hat{\rho})$ and SCR Values for Segmentation Group (SEG_GRP)

Segmentation Group	$\hat{R}(\hat{\rho})$, (S.E.)^	SCR, (S.E.)^
No segmentation group code	0.951 (0.0037)	1.024 (0.0098)
All around average I (homeowner skewed)	0.972 (0.0016)	0.897 (0.0027)
All around average II (renter skewed)	0.962 (0.0033)	0.910 (0.0058)
Economically disadvantaged I (homeowner skewed)	0.963 (0.0024)	0.892 (0.0061)
Economically disadvantaged II (renter skewed)	0.975 (0.0062)	0.831 (0.0283)
Ethnic enclave I (homeowner skewed)	0.968 (0.0029)	0.946 (0.0138)
Ethnic enclave II (renter skewed)	0.968 (0.0123)	0.746 (0.0391)
Young/mobile/singles	0.963 (0.0051)	0.869 (0.0140)
Advantaged homeowners	0.968 (0.0025)	0.926 (0.0047)

Sources: 2010 Census data, American Community Survey interviews and non-interviews from 2007 through 2011
^ S.E. = standard error

Table F-7. $\hat{R}(\hat{\rho})$ and SCR Values for Tenure (TENSHORT)

Tenure	$\hat{R}(\hat{\rho})$, (S.E.)^	SCR, (S.E.)^
Owner-occupied unit	0.971 (0.0014)	0.928 (0.0026)
Renter-occupied unit	0.961 (0.0024)	0.864 (0.0040)

Sources: 2010 Census data, American Community Survey interviews and non-interviews from 2007 through 2011
^ S.E. = standard error

Table F-8. $\hat{R}(\hat{\rho})$ and SCR Values for MAF Tiger Feature Class Code (MTFCC)

MAF Tiger Feature Class Code	$\hat{R}(\hat{\rho})$, (S.E.)^	SCR, (S.E.)^
G2100: Legal American Indian area	0.960 (0.0028)	0.907 (0.0051)
G2120: Hawaiian homeland	0.972 (0.0056)	1.025 (0.0232)
G2130: Alaska Native Village Statistical Area (ANVSA)	0.953 (0.0043)	1.518 (0.0177)
G2140: Oklahoma Tribal Statistical Area (OTSA)	0.984 (0.0013)	0.923 (0.0032)
G2150: State Designated Tribal Statistical Area (SDTSA)	0.973 (0.0020)	0.878 (0.0032)
G2160: Tribal Designated Statistical Area (TDSA)	0.982 (0.0025)	0.907 (0.0096)
G2170: Joint-use Area	0.963 (0.0052)	0.891 (0.0194)
G2200: Alaska Native Regional Corporation (ANRC)	0.978 (0.0026)	0.711 (0.0103)
G2300: Tribal subdivisions	0.949 (0.0034)	0.883 (0.0069)

Sources: 2010 Census data, American Community Survey interviews and non-interviews from 2007 through 2011
^ S.E. = standard error

Table G. Segmentation Group Codes

Segmentation Group (SG)	Percent Occupied Housing Units	Census 2000 Mail Return Rate	Characteristics
0 – CLUSTERNUM is blank	-	-	-
1 – All around average I (homeowner skewed)	35%	77.3%	- 75% owners - 80% non-Hispanic white - largest % of rural tracts - unemployment, poverty, education and mobility levels are close to national averages - skewed towards older persons
2 – All around average II (renter skewed)	16%	74.2%	- more urban and densely populated than SG 1 - above average % of renters and multi-units - skewed towards younger persons
3- Economically Disadvantaged I (homeowner skewed)	6%	66.5%	- 92% of tracts - 49% black - above average % of children - skewed towards older homeowners - higher percentage unemployment, poverty, receiving public assistance, without high school education
4 – Economically Disadvantaged II (renter skewed)	3%	58.0%	- 99.9% of tract are urban - 54% black and 21% hispanic - 81% renters - 1/3 of households speak a language other than english - highest poverty, public assistance, unemployment of all SGs
5 – Ethnic Enclave I (homeowner skewed)	3%	69.8%	- 61% Hispanic - above-average percentage of children - like SG 6 except less linguistic isolation, lower mobility, higher homeownership, fewer asians, less urban, less densely populated - 43% foreign born, 58% of households speak spanish at home
6 – Ethnic Enclave II (renter skewed)	2%	63.6%	- 59% hispanic, 11% Asian - above average % of children - 75% are renters - 34% linguistically isolated - exclusively urban, most densely populated SG, crowded housing - 50% without high school degree
7 – Young/mobile/singles	8%	67.1%	- densely populated and almost exclusively urban - overwhelming majority of households are non-spousal renters in multi-units - skewed to a more educated population - racial and ethnic diversity
8 – Advantaged homeowners	26%	83.2%	- least racially diverse with 85% non-hispanic white - least densely populated - very high percentage of owners, few multi-unit structures, high education, very low levels of poverty and unemployment, low mobility, few non-spousal households

Source: Boone (2008)