Estimating Reliability and Bias from Reinterviews with Application to the 1998 Dress Rehearsal Race Question

Paul P. Biemer
Research Triangle Institute, Research Triangle Park, NC 27709
Henry Woltman
U.S. Census Bureau, Washington, D.C. 20233

Key Words: Nonsampling error; test-retest; validity; mode effects; index of inconsistency

1. Introduction

Reinterview surveys have been used extensively in census evaluations since 1940's. Reinterview surveys are designed with the objective of estimating measurement reliability, measurement bias, or both bias and reliability.   For estimating reliability, a test-retest reinterview design is typically used. In a test-retest reinterview survey, a subset of the questions from the original survey are reasked to form a second set responses to these questions. Estimates of measurement reliability can be computed based upon the patterns of agreement or disagreement between the two responses. There are two key assumptions made for these estimates. One assumes that the error distributions for responses to the interview and reinterview are identical. This assumption implies that the reinterview survey be conducted using the same mode of interview, the same question wording, interviewers with similar training and expertise, and that it interviews the same respondents as interviewed in the original interview.

A key assumption for test-retest reinterview is that measurement errors in the reinterview are independent from the original interview. This assumption usually implies that the reinterview be conducted long enough after the original interview that correlations between the errors due to respondent memory effects are minimized.

A key assumption for “gold-standard” reinterview design is that the responses produced are essentially free of measurement error and, thus, deviations between the interview and reinterview responses are interpreted as errors in the original interview. The gold standard assumption requires that the reinterview use the most experienced and competent interviewers, the most preferred mode of interview, and questions which elicit highly accurate responses. Probing questions may be used to clarify responses. Further, discrepancies between the original interview reinterview responses may be reconciled (see Forsman and Schreiner, 1991, for a discussion of reinterview survey design).

Often for census evaluation programs, reinterview surveys serve a number of research objectives which lead to compromises in the reinterview design. As a consequence, neither the assumptions for test-retest or gold-standard reinterview are met. For example, quite often the evaluation survey is the post-enumeration survey (PES) used for evaluating census coverage. While the census is conducted by mail, self-administered mode, the PES is usually conducted by face to face or telephone with interviewer-assistance. Moreover, since the objective of the PES is evaluation of census coverage error, the data collection procedures employed for the PES are often quite different from those used for the census. Because the PES data are collected by interviewers, the PES data are often considered as a gold standard for evaluating the bias in the census results. However, the PES may use the same questions as used in the Census and there may be no attempt to reconcile discrepancies between the census and reinterview for most responses.   Such design compromises are in conflict with both test-retest and gold standard assumptions Yet both reliability and bias estimates have been reported from the PES data.

Certainly there is a need for both reliability and bias measures for census questionnaires and processes. However, it is not practical or cost effective to conduct separate test-retest and gold standard reinterview surveys in a census evaluation. As we will show in this paper, it is possible to estimate both reliability and bias measures from a one reinterview survey under a set of assumptions that are quite plausible for census evaluations.

We propose a method for estimating response reliability and response bias based upon latent class analysis (LCA) methods. Then we apply the methodology for the estimation of reliability and bias associated with the revised race question that was used in Census 2000 using data from the 1998 Dress Rehearsal. In the next section, we describe this new approach and contrast it with the traditional approach. Then in Section 3, we apply the LCA approach to data on the "mark one or more" race question that was first introduced in the 1998 Dress Rehearsal and was used subsequently in Census 2000.

2. Estimation of Question Quality Measures from Reinterview Surveys

2.1 Traditional Estimates of Reliability and Bias

In this section, we review two measures that are commonly used in evaluating the error in survey questions: the reliability ratio and the response bias. Then we consider classical methods of estimating these components from test-retest and gold standard reinterview surveys.

Let n denote the size of the sample selected for the reinterview survey. To fix the ideas, we assume simple random sampling; however, extensions to complex unequal probability sampling are straightforward. Let A_i denote the census response and let B_i denote the reinterview response to some survey item for person (or unit) i in the evaluation sample. For example, A_i may be an indicator variable for a particular category of race such as Black/African American, where $A_i = 1$ for a census response of Black to the race question and $A_i = 2$ otherwise with an analogous definition for B_i analogously for the reinterview survey. We drop the subscript i in the following since it will be clear from the context when a variable pertains to the unit-level.

Let B denote the true proportion of 1's (for example, the true proportion of Blacks in our previous example) in the population and let p_{A} denote the proportion of 1's in the sample by the original interview response. The bias in the original interview is defined as

$$B_A = \mathrm{E}(p_A) - \mathrm{B}. \qquad (1)$$

Next, consider the usual definition of the reliability ratio from psychometric theory. Using the notation in Biemer and Stokes (1991), we define the reliability ratio, R, for dichotomous variables as the proportion of total variance that is true score variance or, equivalently, 1 minus that proportion of total variance that is error variance. Let Var(A) denote the total variance of an original interview response and note that $\mathrm{Var}(A) = \mathrm{E}_1\mathrm{Var}_2(A)+\mathrm{Var}_1[\mathrm{E}_2(A)]$ where E_1 and Var_1 denote expectation and variance, respectively, with respect to the selection of the evaluation sample and E_2 and Var_2 denote expectation and variance, respectively, conditional on the selected sample. Then Biemer and Stokes (1991) define the reliability ratio as

$$R = \frac{Var_1 E_2(A)}{Var(A)}. \qquad (2)$$

A measure of *un*reliability that is used in many Census Bureau reports is the index of inconsistency, I, defined as 1-R.

Now consider the estimation of B_A and R by traditional methods. Figure 1 shows the usual cross-classification table, denoted by AB, for interview and reinterview for a dichotomous response variables A and B, where A(or B)=1 denotes a positive response and A(or B)=2 for a negative response. Define the "net difference rate" (*NDR*) as

$$NDR = p_{1+} - p_{2+} \qquad (3)$$

Under the assumption that the reinterview survey is the gold standard measure (or true response), Biemer and Stokes show that

$$\mathrm{E}(NDR) = B_A. \qquad (4)$$

That is, *NDR* is an unbiased estimator of the bias in the original interview response.

	***B*=1**	***B*=2**	
***A*=1**	p_{11}	p_{12}	p_{+1}
***A*=2**	p_{21}	p_{22}	p_{+2}
	p_{1+}	p_{2+}	p_{++}

Figure 1. AB (Interview-Reinterview) Cross-Classification Table

To estimate R we assume the first two moments of the error distribution for A and B are equal; i.e., $\mathrm{Var}_2(A) = \mathrm{Var}_2(B)$ and $\mathrm{E}_2(A) = \mathrm{E}_2(B)$. Further, we assume that A and B are conditional independent given the evaluation sample. Under these assumptions, Biemer and Stokes show that

$$\hat{I} = \frac{p_{12} + p_{21}}{p_{1+}p_{+2} + p_{+1}p_{2+}}. \qquad (5)$$

is a consistent estimator of I and, thus,$\hat{R} = 1 - \hat{I}$ is a consistent estimator of R. Another estimator of R is Cohen's kappa statistic (see Cohen, 1960) given by

$$\kappa = \frac{P_0 - P_e}{1 - P_e} \qquad (6)$$

where $P_0 = p_{11}+p_{22}$ and $\mathrm{P}_e = p_{1+}p_{+1} + p_{2+}p_{+2}$. Hess, Singer, and Bushery (1999) show the equivalence between 6 and 1-I so either estimator can be used.

When the assumptions associated with the estimation of B_A. and R are violated, the estimators *NDR* and$\hat{R}$ (or 6) are biased. It is not uncommon for reinterview survey responses to be biased in the same direction as the interview response. In that situation, |NDR| will be smaller in expectation than $|B_A.|$. Further, if the error distributions of A and B are not homogeneous, then $\hat{R}$ estimates the average reliability for both interview and reinterview, i.e., $(R_A+R_B)/2$ where R_A and R_B are the reliabilities for the interview and reinterview, respectively. Thus, if $R_A < R_B$ then $\hat{R}$ will overestimate R and if $R_A > R_B$, then$\hat{R}$ underestimate R.

2.2 Latent Class Model for Estimating Reliability and Bias

An approach which addresses the shortcomings of traditional analysis is a model-based approach using LCA. Like traditional analysis, the LCA method assumes that interview and reinterview errors are independent; however, it makes no assumption that the error distributions for interview and reinterview are equal. This feature is particularly important for census evaluations for the reasons cited in Section 1. Through an appropriate LCA model, the misclassification probabilities associated with interview and reinterview can be estimate separately. Thus, estimates of the bias and variance components for each interview can be computed. In order to do this, the LCA approach makes additional assumptions which are often plausible and testable for census evaluation research.

The model will be described for dichotomous response variables and then extended to the case of polytomous variables in the application. Let *X* denote the true but unobserved (latent) classification for an individual in the sample, where *X* = 1 if the individual is a true positive and *X* = 2 if a true negative. Let B denote P(*X*=1), 2_A and 2_B denote the false negative probability for the interview and reinterview, respectively, and N_A and N_B denote the false positive probability for the interview and reinterview respectively. That is, 2_A = P(*A*=2|*X*=1) and N_A = P(*A*=1|*X*=2) with analogous definitions for *B*. Further, under the assumption that *A* and *B* are conditionally independent given *X*, we can write the expected cell probabilities in Figure 1 in terms of the five parameters: B, 2_A, 2_B,N_A and N_B. Provided the model is identifiable, maximum likelihood estimation can be used to estimate these parameters.

Denote the MLEs of the parameters by the parameter's symbol with a 'hat'. Using the MLE's from the LCA, we can estimate B_A , R_A , B_B , and R_B using the following formulas found in Biemer and Stokes (1991):

1. $\hat{B}_A = p_{1+} - \hat{\pi}$ and $\hat{B}_B = p_{+1} - \hat{\pi}$
2. $\hat{I}_A = \dfrac{\hat{\pi}(1-\hat{\theta}_A)\hat{\phi}_A + (1-\hat{\pi})\hat{\theta}_A(1-\hat{\phi}_A)}{p_{1+}p_{2+}}$ and $\hat{I}_B = \dfrac{\hat{\pi}(1-\hat{\theta}_B)\hat{\phi}_B + (1-\hat{\pi})\hat{\theta}_B(1-\hat{\phi}_B)}{p_{+1}p_{+2}}$
3. $\hat{R}_A = 1 - \hat{I}_A$and $\hat{R}_B = 1 - \hat{I}_B$.

Unfortunately, with five parameters and only four cells in the AB table, the model described above is not identifiable. However, we can employ an device suggested by Hui and Walter (1980) to achieve an identifiable model. Let *G* denote a grouping variable having *K* categories. For example, *G* may denote Hispancity where *G* = 1 for a person reporting Hispanic origin and *G* = 2 for a person reporting non-Hispanic origin. We can extend the LCA model to the GAB table by indexing the parameters defined for the AB table by *g*. For example, for *K*=2 groups, there are 10 parameters: B_g, 2_{Ag}, 2_{Bg},N_{Ag} and N_{Bg}, for *g* = 1,2. However, with only eight cells, the model is still over-parameterized and some restrictions are necessary for identifiability. Hui and Walter (1980) show that an identifiable model with eight parameters results with the assumptions (a) $2_{A1} = 2_{A2} = 2_A$, say, (b) $2_{B1} = 2_{B2} = 2_B$, (c) $N_{A1} = N_{A2} = N_A$,, and (d) $N_{B1} = N_{B2} = N_B$. Adding the four parameters are B_1, B_2, the overall mean, and the proportion of the population in group 1 brings the total number of parameters to eight. Since the model is fully saturated, there no degrees of freedom remaining to assess model fit in the dichotomous variable case. However, all the statistics in (1)-(3) and their standard errors can be still estimated.

This model can be expressed as a hierarchical log-linear model with terms {GX, AX, BX}(see Hagenaars, 1993). Any software that can fit log-linear models with latent variables can be used to obtain the MLE's of the model parameters. The software used in the illustrations to follow is REM (Vermunt, 1997). In the next section, we apply this model to data from the 1998 Dress Rehearsal in order to estimate the reliability and bias for the new census race question.

3. Application to the Race Question Evaluation

Since1977, the Office of Management and Budget (OMB) standard for reporting race characteristics in government surveys specified four basic racial categories: American Indian/Alaskan Native, Asian/Pacific Islander, Black, and White . (See OMB Policy Directive No. 15.) However, shortly after the standards were introduced, they were criticized for not adequately reflecting the increasing racial and ethnic diversity of the population of the United States.

To address these concerns, OMB established the Interagency Committee for the Review of Racial and Ethic Standards in1994. The members of this committee included more than 30 agencies representing the many federal requirements for data on race and ethnicity. The work of this subcommittee culminated in 1997 when OMB announced a revise set of standards (see the *Federal Register* Notice of October 30, 1997 number 62 FR 58782-58790). Under these revised standards, agencies are required to offer individuals the opportunity to select one or more races when reporting information on race in Federal data collections. The five minimum race categories are American Indian/Alaska Native, Asian, Black/African American, Native Hawaiian/Other Pacific Islander, and White.

Census 2000 was the first nationwide implementation of the revised standards. Although the "mark one or more" race question is more complex than the old "mark one"question, it is expected that new question will more accurately capture the increasing diversity of the Nation's population. Results from Census 2000 will display the full range of single and multiple race reporting by the American people. The implementation of the new race question in Census 2000 followed from years of extensive evaluation of the question.

To date, there have been no evaluations that specifically considered the reliability of the new race question. The 1995 National Content Test conducted a test-retest survey for a race question similar to that used in the 1990 Census, but with a single category for "multiracial." This study found no increased in inconsistent reporting with this question. However, "mark one or more" race question was not evaluated in that study.

In remainder of this section, we consider the estimation of reliability and bias from the 1998 Dress Rehearsal data on the "mark one or more" question using the LCA approach described in Section 2. The questionnaire used in the1998 Census Dress Rehearsal is almost identical to the Census 2000 questionnaire. Essentially the same version of the race question was also asked in the Dress Rehearsal PES; this the PES will serve as a reinterview response for the race evaluation.

3.1 The 1998 Dress Rehearsal Data

The 1998 Dress Rehearsal was conducted in 1998 in three sites: Columbia, S.C. and eleven surrounding counties (referred to as Rural S.C.), Menominee County, WI, and Sacramento, CA. Since the Menominee County site was not part of the race question evaluation, our analysis will focus on three on sites: Columbia, Rural S.C., and Sacramento. Sacramento used an Integrated Coverage Measurement (ICM) methodology while a Post Enumeration Survey (PES) methodology was used in the two S.C. sites. Without going into the differences between these two data collection approaches, they were essentially the same in their implementation of the race question and will be treated as such in our analysis. We will refer to the coverage evaluation surveys in both sites as the Dress Rehearsal PES.

The Dress Rehearsal PES is a reinterview that was conducted in a sample of census blocks by enumerators who were specially trained in the PES data collection operations. In addition to collecting the names of the current residents of the sample households, the PES also collected remeasurements of the basic demographic information for those residents. The data in our analysis is for all persons who were matched between the census and the PES and who answered the race questions in both. The total sample size for the evaluation is 40,519 persons which was split between Columbia, Rural S.C., and Sacramento with sample sizes 14,273, 12,711, and 13,535 persons, respectively. In our analysis, a person is classified as Hispanic if they were reported as Hispanic in either the census or the PES. The combined sample contains 5,161 persons classified as Hispanic and 35,358 persons classified as non-Hispanics.

In the next section, a LCA approach such as that described in Section 2 will be applied to the race data and estimates of reliability and bias for each race with be obtained. For race reports, the concept of bias is somewhat vague since it implies the existence of a true race classification for an individual. Since race is a somewhat subjective concept, a "true race" may not be meaningful for all individuals. However, one can still conceptualize of a preferred method of obtaining race data - one that is devoid of influences that would cause instability in responses to the race question. Deviations from this preferred response may then be interpreted as a type of bias. The advantage of such a concept is that it allows an examination of the systematic errors in the determination of race that may be related to the mode of interview. Thus, this is the concept classification error bias will be used in the remainder of this section.

3.2 Race Analysis for the 1998 Dress Rehearsal Data

Sample selection for the PES was not an equal probability sample of Dress Rehearsal persons; however, simple random sampling will be assumed in our analysis for simplicity. The unequal weighting effects in the sample are quite small, however, which suggests a weighted analysis should not differ much from our unweighted analysis.

In analogy to the LCA model described in Section 2.2, let *A* denote the Dress Rehearsal census classification and let *B* denote the PES classification of an individual's race. For the grouping variable *G*, we chose two variables: *S* denoting the site where *S* = 1 for Sacramento, *S* = 2 for Rural S.C., and *S* = 3 for Columbia and *H* denoting Hispanic (*H*=1) and non-Hispanic (*H*=2) origin classifications.

The model for the analysis is an extension of the model described in Section 2.2 to incorporate two grouping variables, *S* and *H*. In hierarchical log linear model notation, the model we will use is {SHX, AHX, BHX, AS, BS}. Other models were explore in our analysis; however, this model provided the best fit. The SHX term in the model specifies different race prevalence rates for all six site by Hispancity combinations. The terms AHX, and BHX indicate that the race classification error rates for both the census and the PES vary by Hispancity. To achieve an identifiable model, the ASX and BSX terms are set to zero; this means that the false positive and false negative error rates are assumed not to vary independently across sites. However, the presence of the AS and BS terms allow some *dependent* variation of these rates by site. Further the absence of terms involving the AB-interaction is consistent with the assumption of independent classification errors.

Rather than consider each race category as a dichotomous variable, the dependent variables in our analysis are five-category race response variables, *A* and *B* where *A*,*B*=1,...,5 correspond to White, Black, API, Some Other Race, and More Than One Race, respectively. The "Some Other Race" category contains all persons who marked any single race category other than White, Black, and API or wrote-in a single other race. The "More Than One Race" category contains all persons who marked two or more race categories or one category and wrote-in one or more other categories. API is formed by collapsing Asian, Native Hawaiian, and other Pacific Islander categories. Since there are 150 cells in the SHAB table and 126 parameters in the model, the model was fit with 24 degrees of freedom to test the fit of the model. This model was rejected using the standard chi-square test criterion. However, with over 40,000 observations the power of the standard test is approximately 1.00 and is therefore not suitable for assessing model fit. A better indicator of model fit is provided by the index of dissimilarity, *d*, which is the proportion of observations that would be misclassified under the model. For this model, *d* was less than 0.4 percent, indicating excellent agreement between model and data and a well-fitting model.

The LCA provided estimates of B_g, 2_{Ag}, 2_{Bg},N_{Ag} and N_{Bg}, for the six site by Hispanicity groups. These estimates were then used to estimate R_A, R_B, B_A and B_B using the estimation formulae in Section 2.2. The results of these analyses are given in Tables 1a, 1b, 2a, and 2b.

For each race category, Tables 1a and 1b presents the unweighted race prevalences and the *NDR* for the census and the PES as well as the LCA estimate of the overall reliability $R = (R_A + R_B)/2$ by Hispancity and site. Note that the estimate of *R* in Table 1a is comparable to 6 (or equivalently 1-*I*) discussed previously since both are consistent estimators of the average of the reliabilities for measures *A* and *B*. However, one advantage of the LCA estimator of *R* is that it is never negative whereas 6 is sometimes negative. A disadvantage of the LCA estimator is that its standard error maybe somewhat larger than the standard error of 6. A comparison of standard errors of the estimators is not within the scope of the present paper, however.

Considering the overall results (first panel of Table 1a), we see that the Census and PES estimates of race prevalence are significantly different for three race categories: Black, Some Other Race, and More Than One. However, as can be seen from the other second and third panels, the difference is primarily due to large discrepancies among the Hispanics. Also, from Table 1b we see that the differences in Sacramento and Columbia are much larger differences than those is Rural S.C.

If we assume the PES is the gold standard for the classification of race, the census-PES differences can be interpreted as indicators of measurement error bias. This interpretation would lead to the conclusion that the Census tends to overestimate persons of multiple races by about 3 percentage points and underestimates persons of Some Other Race by about 3 percentage points. In addition, there is a slight (about 1 percentage point) bias in the Black race. Although significant, the biases for the non-Hispanic race classifications are small while the biases for Hispanics are quite large for all categories other than API.

Next consider the reliability of the census under the usual test-retest assumptions, which as we mentioned, probably do not hold for the PES. From Tables 1a and 1b, we see that White and Black race are reported with fairly good reliability (R>0.90) both for the total population overall and for non-Hispanics. However, for Hispanics, all races display poor reliability. Interestingly, Some Other Race appears to be reported more reliably by Hispanics than by non-Hispanics.

If we are not willing to make the assumptions associated with traditional analysis, the LCA model estimates of bias and reliability may be more appropriate. The LCA estimates of R_A, R_B, B_A and B_B are shown in Tables 2a and 2b. Note that the estimates in tables satisfy $(R_A + R_B)/2 = R$ and $B_A - B_B = NDR$ in Tables 1a and 1b.

The equality of R_A and R_B can be tested via a likelihood ratio test. The restricted model sets the conditional probabilities P(A|SHX) equal to P(B|SHX) and the unrestricted model removes this restriction. The restricted model was rejected with p<0.001 and, hence, the hypothesis $R_A =$

R_B must also be rejected. This suggested that the assumption of homogeneous error distributions made for classical test-retest analysis does not hold for these data and the assumption of B_A and B_B must be rejected as well.

The largest differences between R_A and R_B in these tables occurs for the More Than One Race and Some Other Race categories. Note that in some cases, census reliability is greater than PES reliability. The estimates of B_A and B_B in the tables suggest that the PES race classification is also biased, in some cases, more so that the census. Further, the bias is often in the same direction for both surveys leading to underestimates of the bias in the Census. For this reason, the bias estimates from the LCA are often considerably larger than the bias estimates from the traditional gold standard analysis.

The estimates in Tables 2a and 2b provide a much different picture of reliability and bias than the estimates in Tables 1a and 1b, the latter tables arising from classical analysis assumptions. This suggests that the LCA can provide very different insights into the quality of both the census and the PES data.

.

4. Summary and Conclusions

This paper considers the problem of estimating measurement reliability and bias from reinterview data. The traditional test-retest estimator of response reliability assumes that the classification error probabilities for interview and reinterview are identical and that classification errors are independent. These assumptions are seldom satisfied in practice, particularly for census reinterview evaluation studies since in a census, the original response is obtained by mail/self-administration and the evaluation reinterview response is usually obtained by face to face or telephone interviewer assisted administration. Mode differences tend to invalidate the assumptions made for traditional analysis. We propose a new method for estimating measurement reliability and bias based upon a LCA model that relaxes the assumptions of classical reinterview analysis and which may be more appropriate for census evaluation. This method allows the estimation of reliability and bias separately for the original and reinterview. We applied the methodology for the estimation of reliability and bias associated with the revised race question that was used in Census 2000 using data from the 1998 Dress Rehearsal. Our analysis provides evidence that the classical assumptions are not satisfied for these data and that the LCA estimates may be better measures of reliabilities and biases associated with the classification by race in the census and the PES.

REFERENCES

Biemer,P., and Stokes, L. (1991). Approaches to Modeling Measurement Error. In P.P. Biemer, R. Groves, L. Lyberg, N. Mathiowetz, and S. Sudman (Eds.), *Measurement Errors in Surveys*. New York: John Wiley & Sons.

Biemer, P., Bushery, J., and Flanagan, P. (1997). "An Application of Latent Markov Models to the CPS," Internal U.S. Census Bureau Technical Report.

Cohen, J. (1960). "A Coefficient of Agreement for Nominal Scales," *Educational and Psychological Measurements*, 20, 37-46.

Forsman, G. and Schreiner, I. (1991). "The Design and Analysis of Reinterview: An Overview," in P.P. Biemer, et al. (eds.), *Measurement Errors in Surveys*, John Wiley & Sons, NY.

Hui, S.L. and S.D. Walter (1980). "Estimating the Error Rates of Diagnostic Tests," *Biometrics*, 36, 167-171.

Hagenaars, J. (1993). *Loglinear Models with Latent Variables*, Sage University Paper Series, Quantitative Applications in the Social Sciences, 07-094, Newbury Park, CA: Sage.

Hess, J., Singer, E., and Bushery, J. (1999). "Predicting Test-Retest Reliability from Behavior Coding," *International Journal of Public Opinion Research*, Vol. 11, No. 4, pp. 346-360.

Vermunt, J. (1997). *REM: A General Program for the Analysis of Categorical Data*, Tilburg, University.

Table 1a. Initial Race Response and Net Difference Rate for Non-Hispanics, Hispanics and Overall

	Overall				Non-Hispanics				Hispanics			
Race	P_1	P_2	*NDR*	*R*	P_1	P_2	*NDR*	*R*	P_1	P_2	*NDR*	*R*
White	55.2	54.8	0.4	96.4	59.8	58.5	1.3^{***}	98.8	23.3	28.8	-5.5^{***}	50.8
Black	27.1	28.4	-1.3^{***}	93.8	30.5	30.0	0.5	97.5	3.5	17.0	-13.5^{***}	44.4
API	5.9	5.7	0.2	80.0	6.4	6.3	0.1	85.8	2.2	1.6	0.7	26.1
Some Other Race	5.1	7.8	-2.7^{***}	56.9	1.5	2.5	-1.0^{***}	31.6	29.8	44.3	-14.6^{***}	77.5
More Than One	6.8	3.4	3.4^{***}	28.1	1.8	2.7	-0.9^{**}	35.1	41.3	8.4	32.9^{***}	45.1

*significant at = 0.05; **significant at = 0.01; ***significant at = 0.001

Table 1b. Initial Race Response and Net Difference Rate by Site

	Sacramento				Rural SC				Columbia, SC			
Race	P_1	P_2	*NDR*	*R*	P_1	P_2	*NDR*	*R*	P_1	P_2	*NDR*	*R*
White	52.0	46.4	5.7^{***}	93.6	61.8	62.7	-0.9	98.7	52.2	55.7	-3.5^{***}	96.6
Black	12.9	12.0	1.0^{*}	90.4	34.3	34.8	-0.5	97.8	34.1	38.2	-4.1^{***}	92.4
API	15.9	15.5	0.4	83.2	0.6	0.5	0.1	66.4	1.1	1.0	0.1	67.6
Some Other Race	13.7	19.4	-5.8^{***}	66.6	0.7	1.0	-0.4	39.9	0.9	2.8	-1.9^{***}	12.5
More Than One	5.5	6.8	-1.3^{**}	31.1	2.7	1.1	1.6^{**}	26.5	11.7	2.3	9.4^{***}	29.1

*significant at = 0.05; **significant at = 0.01; ***significant at = 0.001

Table 2a. Reliability and Bias for the Census and PES by Race and Hispanicity

	Overall				Non-Hispanics				Hispanics			
Race	R_A	R_B	B_A	B_B	R_A	R_B	B_A	B_B	R_A	R_B	B_A	B_B
White	97.0	95.9	2.7^{***}	2.3^{***}	98.9	98.6	1.6^{***}	0.4	50.1	51.4	9.7^{***}	15.2^{***}
Black	92.3	95.2	-2.1^{***}	-0.8^{**}	98.3	96.8	0.1	-0.4	12.2	76.6	-17.0^{***}	-3.5^{***}
API	80.0	79.9	-1.0^{***}	-1.2^{***}	84.1	85.6	-0.8^{*}	-1.0^{**}	34.8	17.4	-2.2^{**}	-2.9^{***}
Some Other Race	47.2	66.6	-1.8^{***}	0.9^{**}	16.1	47.5	-0.3	0.7^{*}	70.2	84.8	-12.3^{***}	2.3^{**}
More Than One	41.7	14.5	2.3^{***}	-1.1^{***}	39.2	31.0	-0.5	0.4	73.4	16.7	21.9^{***}	-11.1^{***}

*significant at ∓ 0.05; **significant at ∓ 0.01; ***significant at ∓ 0.001

Table 2b. Reliability and Bias for the Census and PES by Race and Site

	Sacramento				Rural SC				Columbia, SC			
Race	R_A	R_B	B_A	B_B	R_A	R_B	B_A	B_B	R_A	R_B	B_A	B_B
White	92.9	94.4	7.2^{***}	1.5^{**}	99.3	98.2	0.9	1.8^{***}	98.5	94.7	0.0	3.4^{***}
Black	90.8	90.0	1.3^{**}	0.3	98.0	97.5	-0.9	-0.5	89.4	95.2	-6.4^{***}	-2.3^{***}
API	83.7	82.7	-3.6^{***}	-4.0^{***}	57.9	74.9	0.2	0.1	65.5	69.7	0.3	0.2
Some Other Race	54.6	78.6	-6.5^{***}	-0.7	34.3	45.5	0.2	0.6	14.5	10.6	0.8	2.6^{***}
More Than One	28.5	33.6	1.5^{**}	2.9^{***}	45.3	7.7	-0.3	-2.0^{***}	52.1	6.1	5.4^{***}	-4.0^{***}

*significant at ∓ 0.05; **significant at ∓ 0.01; ***significant at ∓ 0.001