THE USE OF STANDARDS IN SURVEY ESTIMATION

D. G. Horvitz, R. E. Folsom, and L. M. Lavange
Research Triangle Institute
D. G. Horvitz, P. O. Box 12194, Research Triangle Park, N.C. 27709

INTRODUCTION

This paper argues the need for establishing standards for evaluating the contribution of the systematic components of each of the various sources of nonsampling errors in sample surveys. The use of variances and standard errors of estimates computed from survey data to assess the reliability of survey statistics has increased markedly in recent years, although still not enough to be considered a routine practice. On the other hand, the reporting of the level of systematic error, or bias, in estimates computed from survey data is almost nonexistent. Yet, in the relatively few instances in which accurate estimates of the bias in survey statistics have been reported, such estimates have been far from negligible, and usually have dominated the sampling error in those same statistics.

Surveys which require respondents to report events that occurred at some time in the past to sample persons are subject to rather significant systematic coverage and measurement errors. The realized level of error (net bias) may be related to one or more controllable design factors, such as choice of frame, mode of interview, number of proxy respondents, length of the reference period, and the use or non-use of temporal bounds. The paper proposes, just as is expected for the sampling error in survey estimates, that the nonsampling bias in those same estimates be routinely assessed through (i) the use of specific design factor levels as "standards of accuracy," (ii) collecting the survey data at the chosen set of standard design factor levels for at least a subsample, preferably randomized, and (iii) estimating the systematic error effects or biases relative to the chosen standards at the alternative (non-standard) design factor levels occurring in the survey. The paper also proposes that survey statistics be routinely adjusted for measurement biases based on the chosen standards, just as they are now routinely adjusted to reduce coverage and nonresponse biases.

An accuracy standard for measurements is never absolute. A given standard, in a sense, represents the concensus "best" level to use for a given survey measurement factor to obtain the data required at the current state of the art. For example, with respect to the respondent rule (a controllable design factor) in interview surveys, it is generally accepted that the data collected from self-respondents (first level) about events occurring to themselves in the past are more accurate than the same data obtained from proxy respondents (second level). In order to be able to evaluate or otherwise interpret estimates of nonsampling biases relative to a set of standards of accuracy, it is essential, for consistency, that the same set of standards be used across all surveys. Hence, the paper further proposes a procedure for the selection and publication of a single set of standards of accuracy for use by all survey practitioners in estimating nonsampling biases in survey estimates.

STANDARDS OF ACCURACY IN SURVEY ESTIMATION

The use of standards of accuracy in survey estimation is implicit in several procedures currently practiced by survey statisticians. For example, whenever the most current census estimate of the age, race, sex distribution of the U. S. household population is used for a post stratification adjustment, the census data are being accepted as a standard of accuracy. Thus, in large samples, at least, the coverage bias is claimed to have been reduced by the post stratification procedure. The amount of the reduction, or level of coverage bias in the unadjusted data relative to the chosen standard, can be estimated by comparing the sample estimates obtained with and without the post stratification.

Similarly, whenever the basic set of sample inclusion probabilities is adjusted for unit nonresponse using weighting classes, the data collected from the respondents belonging to these weighting classes are being accepted as an estimation standard. Hence, the net sampling bias due to nonresponse in the unadjusted survey data, at least relative to this standard, can be estimated for each survey variable by comparing the sample estimate obtained with the unadjusted sample weights to that obtained with adjusted weights. Similarly, the use of hot deck imputation implies acceptance of the data in the donor group as a standard of accuracy. Again, the net bias due to item nonresponse for a specific survey variable relative to the donor group standard can be estimated by comparing the estimates obtained with and without imputation.

In our opinion, the survey research community should not continue to ignore the bias in the measurement process. Just as adjustments are made to reduce coverage, unit nonresponse and item nonresponse biases in survey estimates, so is it possible, with appropriate survey designs and choice of accuracy standards, to use the data collected to adjust substantive survey estimates for measurement bias. For example, the Health Interview Survey (HIS) hospital discharge estimates can be adjusted for recency bias because of the use of overlapping reference periods in the survey design. If the reporting of events which occurred in the two weeks prior to interview is accepted as more accurate than the reporting of events which occurred earlier, then the overlapping reference period design permits "standard unbiased" estimates to be derived using all of the sample data. These design based adjusted estimates should have no more actual bias than estimates derived from

events reported for the two weeks prior to interview.

Rotating panel designs, such as used in the National Crime Survey (NCS) and in the Survey of Income and Program Participation (SIPP), also permit "standard unbiased" estimates of biases arising from unbounded measurements and time-in-panel, as well as from he length of the time interval between the interview and the occurrence of events of interest. The term "standard unbiased," as implied above, is used to refer to accuracy relative to a standard measure, which, in absolute terms could still be biased.

Numerous factors contribute to the net bias due to systematic errors introduce by the measurement process. For example, a specific measurement design might specify the following factors and a specific level for each:

FACTOR	LEVELS
Mode of interview	Personal, telephone, mail
Respondent rule	Self, proxy
Length of recall	One month, two months, three months, etc.
Type of recall	Bounded, unbounded
Interview method	Paper and pencil, computer assisted
Administration	Self, by interviewer
Time-in-sample	1st, 2nd, etc. (For longitudinal surveys)

With the choice of a suitable accuracy standard, the net bias associated with a specific measurement design factor and level is estimable, relative to the chosen standard, for each substantive variable/item/event measured in a given survey. For example, estimation of the measurement bias arising through the use of proxy respondents in a given survey implies self-response and proxy measurement in that same survey of a consistent probability subsample of cases. The best design would reverse the order of the two types of respondent (self- and proxy-) in half of the subsample. The differences between the self- and proxy- statistics for the subsample provide unbiased estimates of the measurement biases associated with the use of proxy respondents.

Similarly, through the use of agreed upon standards, such as personal interview, one month recall, bounded recall, etc. the relative bias in each of the other measurement design factors can be individually assessed. It is also useful to consider the net bias associated with a particular measurement design. For example, an agreed upon measurement design standard might be:

Personal interview
Computer assisted
Self-respondent
One-month bounded recall

The actual measurement design used in a given survey could depart from this standard in a number of ways. It could include, for example, the use of paper and pencil, face-to-face and telephone interviews, with proxy- as well as self-respondents, together with unbounded six-month recall. However, it is not essential to estimate reliably the bias associated with each of these measurement design factors at the level used in the specific survey. Rather, through measurement of a suitable subsample of cases using the measurement design standard, an unbiased estimate of the net bias associated with the actual measurement design used can be obtained.

Having agreed upon a "measurement design standard" the most appropriate allocation of survey cases as between the more expensive standard measurement and the less expensive, less accurate "actual measurement design" remains to be determined. Cost and error models are needed for design optimization, models which properly reflect the trade-off between allocating resources to reduce sampling error and those concerned with estimating reliably the net bias in the "actual measurement design" relative to the "measurement design standard."

ESTABLISHING STANDARDS

The survey research community can benefit through the adoption and use of a single set of accuracy standards for controllable measurement design factors. It is proposed that responsibility for establishing standards reside in the Survey Research Methods Section of the American Statistical Association. A Standards Committee, with rotating membership appointed by the Section Chair, should periodically publish measurement design standards in several statistical journals such as The American Statistician, Survey Methodology, and International Statistical Review.

It is recognized that a given standard may not apply to all the variables measured in surveys. While there could be general standards which apply to all variables, or at least those in a particular category, there could also be very specific standards which apply to particular variables only. An example of a general respondent rule standard applicable to all variables is "self-interviews for all persons 16 years and older." On the other hand, while "personal interview" might be established initially as a general mode-of-interview standard for all variables, it is possible that other interview modes may eventually prove to be more accurate for some variables. It is expected that only general standards would be established initially.

It should not be expected that a given set of measurement standards for surveys is without bias. Nor should we expect that the quality of measurements will remain constant over time. Having chosen a set of measurement standards, those concerned with improving the quality of survey measurements should undertake research aimed at determining ways to improve upon the standards. The direction of such research will clearly be guided by the accumulation of information on measurement biases based on the chosen set of standards. The standards should be modified and refined as breakthroughs in achieving higher quality measures are realized from the methodological research.

ESTIMATING SYSTEMATIC ERROR LEVELS

Given a set of accuracy standards, the net bias in the measurement design and in the measurement design components used in each and every sample survey should be routinely

estimated, relative to the chosen standards. This implies, as suggested above, a willingness to collect data for a design consistent probability subsample using the "standard measurement design". The difference between the estimate obtained for a similar subsample using the "actual measurement design" and the estimate using the "standard measurement design" is a design-based standard unbiased estimate of the net bias generated by the "actual measurement design". Design-based standard unbiased estimates of the realized measurement bias in National Crime Survey (NCS) statistics relative to the statistics generated by specific measurement standards are presented below.

It is not always possible to collect the data of interest using a "standard measurement design" on a strict probability subsample. Standard unbiased estimates of the realized measurement design bias in strict terms are not possible in this situation. Still, reasonable estimates of the net measurement bias relative to the standard may be possible using multivariate regression provided there is sufficient balancing of the total sample across the eligible levels of the measurement factors included in the measurement design. With adequate balance relative to each potential confounder, the resulting level of confounding in the estimated differences between levels for a given measurement factor should be minimal.

NATIONAL CRIME SURVEY EXAMPLES

Among the measurement design factors that contribute to the bias in NCS estimates are mode of interview, respondent rule, length of recall, type of recall, and time-in-sample. An in-depth study (Lavange amd Folsom, 1985) to develop statistical methods appropriate for estimating and adjusting for biases due to systematic errors in the NCS data was recently undertaken at Research Triangle Institute. Although focussed primarily on various modelling approaches to this problem, the results provide illustrations of the use of standards for survey measurement processes and their impact on the resulting estimates.

The NCS design consists of a stratified multistage cluster sample of approximately 60,000 dwelling units that are interviewed every six months for three and one-half years for a total of seven interviews. The sample is randomly allocated to six rotation groups and each rotation group is further divided into six panels corresponding to months of interview. A new rotation group is rotated into the sample each month. Data collected during this initial interview are used to bound subsequent interviews and are not included in reported estimates. In each interview month seven rotation groups are interviewed, one of which corresponds to a bounding interview group. Respondents are asked to report victimizations occurring during the past six months. Thus, for each month in a reference period, incidents are reported from six panels and seven rotation groups.

The NCS rotating panel design just described permits unbiased estimation of the effects of systematic errors due to three factors randomized in the design, namely, errors due to unbounded first interviews (forward telescoping), time-in-sample for the dwelling unit (conditioning effect), and time lag between interview month and month of victimization occurrence (recency effect). Two variables were defined for use in measuring these effects:

1) Time-in-sample is the total number of times an NCS interview was conducted at a dwelling unit including the current interview.

2) Recency is the time in months from the reported victimization occurrence to the time of interview.

Time-in-sample ranged from one to seven with the first time in sample corresponding to the initial bounding interview. Recency ranged from one to six. In order to estimate systematic measurement error components for these two variables, a subset of the NCS longitudinal file, inclusive of interviews contributing to the reference period of January through June 1978, was analyzed.

In Table 1, victimizations for personal crimes with contact and personal crimes without contact (larcenies) reported as occurring in the first two months prior to interview are assumed to be free of systematic reporting errors. Thus, the accuracy standard for the recency variable for this illustration is the combined victimization rate for recency values 1 and 2. The table provides unbiased estimates of the bias, relative to the chosen standard, in the victimization rates computed from data reported with recency values 3, 4, 5, and 6. Respondent reports in the NCS are significantly lower, than for the standard, for those victimizations which occurred in the fourth, fifth and six months prior to interview.

Table 2 gives unbiased estimates of the bias due to time-in-sample, for personal crimes with and without contact, under the assumption that the second time-in-sample is the best (least biased) standard. If this is a reasonably valid assumption, the estimates in Table 2 indicate that time-in-sample conditioning significantly reduces the reporting of victimizations in the NCS for personal crimes with contact after the third round of interviews.

The difference between the observed victimization rates for the first and second time-in-sample (again assuming the bounded second interview produces the least biased standard) estimates the bias in the unbounded first interview to be approximately seven victimizations more per 1000 person years for contact crimes and 28 more for noncontact crimes.

Initial models were fit to the contact and noncontact personal crime victimization rates that included the main effects of the three design factors and their interactions. RTI's survey regression software package, SURREGR (Holt, 1977), was used to consistently estimate the finite population model parameters and their variance covariance matrix. Recency and time-in-sample (including bounding) were found to interact significantly for personal crimes with contact, but not for personal crimes

without contact. In order to produce efficient, smoothed model based adjusted rates, reduced models were fit to the data in which orthogonal polynomial trends were used to characterize the effects. Weighted least squares methods appropriate for complex survey data (Koch, Freeman, Freeman, 1975) were employed to fit polynomial trend models to the 42 victimization rates defined by the cross-classification of the measurement design factors.

To quantify the overall or net measurement bias of the NCS design due to the occurrence of interviews with less than optimal time-in-sample, bounding and recency factor levels, a measurement design standard was specified. The selected standard, chosen here for illustrative purposes only, consists of victimizations reported during the second, bounded interview as occurring in the first or second month prior to interview. The reduced polynomial trend models were evaluated at the second time-in-sample and first two months of recency in order to produce rates adjusted to this standard. The adjusted and unadjusted rates are given in Table 3 for both types of crimes. In both cases, the net adjustment effect was to increase the reported rate. The rate for contact crimes increased from 37 to 51 victimizations per 1000 person years while the rate for noncontact crimes increased from 94 to 116 victimizations per 1000 person years.

The mean square errors for the unadjusted rates in Table 3 were computed as follows:

$$MSE(R_u) = [Bias(R_u)]^2 + Var(R_u)$$

$$= (R_u - R_a)^2 - Var(R_u - R_a) + Var(R_u)$$

where R_u and R_a denote the unadjusted and adjusted rates respectively. This equality assumes that R_a corresponds to the "true" rate. Comparing the standard error of R_a to the root MSE of R_u in Table 3 indicates that large gains in efficiency were achieved by the adjustment procedure, conditional, of course, on the validity of the assumed accuracy standard.

In the second phase of the NCS analysis, models were fit to a 23 month time series of personal victimization rates that included joint effects of known NCS measurement error sources and important demographic and socioeconomic correlates of victimization. Estimated model parameters were then used to produce robust, time smoothed rates adjusted to a selected standard for each of the survey measurement factors included in the model. The following standards, generally thought to represent the "least biased" levels of the measurement design factors, were selected:

--self-response interviews
--interviews bounded at the person level
--2nd time in panel for respondent
--non-household respondent interviews
--personal (not telephone) interviews
--person responds in every round

In order to determine a least biased level for recency, models were fit to recency that included effects of foreward telescoping and memory loss. Results cited in the literature as well as results of this modelling effort indicated that a recency distribution that weighted the rates associated with a one to three month recall one and one-half to two times higher than rates associated with a four to six month recall was justified. The weighting ratio of 1.75:1 was somewhat arbitrarily chosen in this interval. An average distribution was assumed for all other variables included in the model.

Table 4 gives unadjusted and adjusted rates for personal crimes with contact. Age, sex, and race/ethnicity subgroup specific rates are presented along with their associated measures of variability. In most cases, the adjustment process preserved the differences among demographic subgroups that were detected with the unadjusted rates. This was not the case with respect to race, however. The difference between blacks and whites in contact crimes was reversed in the adjusted rates (a difference of 12 unadjusted and -11 adjusted).

---

[1] It is quite likely that there is a better accuracy standard for recency. The standard used in Table 1 was selected purely for illustrative purposes.

REFERENCES

Holt, M. M.)1077). "SURREGER: Standard Errors of Regression Coefficients from Sample Survey Data," (Revised April 1982 by B. V. Shah). Research Triangle Institute, Research Triangle Park, N.C.

Koch, G. G., Freeman, D. H., Jr., and Freeman, J. L. (1975). "Strategies in the Multivariate Analysis of Data from Complex Surveys." International Statistical Review, Vol. 43, pp. 59-78.

LaVange, L. M. and Folsom, R. E. (1985). "Regression Estimates of National Crime Survey Operations Effects: Adjustments for Nonsampling Bias." Proceedings of the American Statistical Association, Section on Social Statistics, 109-114.

Table 1

Estimates of Net Recency Bias[1] in Annual Victimizations
Reported to have Occurred in Months 3 to 6 Prior to Interview
(Per 1,000 Persons 12 or More Years Old)

	Personal Crimes With Contact		Personal Crimes Without Contact	
Months Prior Interview	Net Bias (Number)	Relative Bias[2] (Percent)	Net Bias (Number)	Relative Bias[3] to (Percent)
3	-14.86 (2.24)[4]	-29.2	-37.33 (4.36)	-32.1
4	-21.56 (2.76)	-42.4	-42.78 (4.27)	-36.7
5	-25.05 (2.74)	-49.3	-52.82 (3.93)	-45.4
6	-29.54 (2.94)	-58.1	-67.47 (3.90)	-58.0

[1]Assumes rates for victimizations reported to have occurred in months 1 and 2 prior to interviews are unbiased.

[2]Relative to overall annual adjusted rate of 50.83 victimizations per 1,000 persons 12 or more years old.

[3]Relative to overall annual adjusted rate of 116.41 victimizations per 1,000 persons 12 or more years old.

[4]Standard errors are shown in parentheses.

Source Data: National Crime Survey, 1978.

Table 2

Estimates of Net Time-in-Sample Bias[1] in Annual Victimizations
(Per 1,000 Persons 12 or More Years Old)

	Personal Crimes With Contact		Personal Crimes Without Contact	
Time-in-Sample	Net Bias (Number)	Relative Bias[2] (Percent)	Net Bias (Number)	Relative Bias[3] (Percent)
3rd	-0.08 (3.23)[4]	- 0.2	-6.23 (5.20)	-5.4
4th	-5.61 (3.60	-11.0	-6.11 (6.06)	-5.2
5th	-9.04 (3.48)	-17.8	-7.54 (5.04)	-6.5
6th	-6.13 (3.37)	-12.1	-1.01 (5.22)	-0.87
7th	-7.14 (3.24	-14.0	-8.23 (4.82)	-7.06

[1]Assumes rates for victimizations reported in second time-in-sample are unbiased.

[2]Relative to overall annual adjusted rate of 50.83 victimizations per 1,000 persons 12 or more years old.

[3]Relative to overall annual adjusted rate of 116.41 victimizations per 1,000 persons 12 or more years old.

[4]Standard errors are shown in parentheses.

Source Data: National Crime Survey, 1978.

Table 3

Unadjusted and Adjusted Annual Victimization Rates for Personal Crimes With Contact and Personal Crimes Without Contact (Per 1,000 Persons 12 or More Years Old)

	Personal Crimes With Contact	Personal Crimes Without Contact
Unadjusted Rate	36.91	94.13
Standard Error	1.46	2.10
Root Mean Square Error	13.88	22.07
Adjusted Rate[1]	50.83	116.41
Standard Error	1.95	4.20

[1]Assumes rates for victimizations reported during the second (bounded) interview as occurring in the first or second month prior to interview are unbiased.

Source Date: National Crime Survey, 1978.

Table 4

Adjusted[1] and Unadjusted[2] Victimization Rates for Personal Crimes With Contact by Age, Sex, and Race/Ethnicity (Per 1,000 Persons 12 or More Years Old)

	Adjusted Rates[1]		Unadjusted Rates[2]	
	$\hat{R}_a$	S.E.($\hat{R}_a$)	$\hat{R}_u$	S.E.($\hat{R}_u$)
1. Age				
12-15 years	74.96	10.01	58.80	4.11
16-19 years	85.34	6.48	71.83	4.38
20-24 years	56.09	6.89	71.78	3.14
25-34 years	41.54	4.16	42.83	2.25
35-49 years	40.32	4.26	22.43	1.27
50-64 years	37.32	3.46	15.31	1.21
65+ years	35.71	3.62	10.80	0.93
2. Sex				
Male	61.00	3.60	48.36	1.67
Female	36.66	3.33	26.29	1.44
3. Race/Ethnicity				
Hispanic	38.19	7.74	42.86	5.01
Black, non-Hispanic	39.62	5.46	46.76	3.57
White, non-Hispanic	50.61	3.30	35.25	1.38

[1]Rates were adjusted to a measurement design standard consisting of 1978 bounded, personal, self-response interviews at the second time-in panel for non-household respondents. A "least biased" distribution for recency and average distributions for all remaining variables in the model were assumed.

[2]Unadjusted rates were computed for 1978 excluding dwelling unit bounding interviews.

Source Data: National Crime Survey, 1978 and 1979.