NATIONAL ASSESSMENT OF EDUCATIONAL PROGRESS (NAEP): NONRESPONSE STUDY

Douglas Wright and Michael P. Cohen, National Center for Education Statistics*
Michael P. Cohen, 555 New Jersey Avenue NW, Washington, DC 20208-5654

KEY WORDS: Bias, variance, simulation, schools, states

Introduction

Recently, the National Assessment of Educational Progress (NAEP) expanded from a national survey to state representative samples on a trial basis. Cooperation generally has been quite good with 38 states participating in the 8th grade mathematics assessment in the first year (1990) and 42 states in 1992.

School cooperation rates with states have varied from state to state from a low of 62% (in 1992) to 100%. The lower response rates have raised concerns about potential nonresponse bias. In 1992 NCES undertook to study the impact of this nonresponse through two simulation projects.

The first project looked at State assessments that were similar to NAEP tests for states with low NAEP response rates. Often states will conduct their own assessments on a census of their schools. For such states we could calculate the difference between the estimated average state assessment score based on the NAEP sample respondent schools and weights and that based on the complete census. We could analyze these differences both at the state level and for substate categories.

The second project took states with 100% response rates and simulated the levels and patterns of nonresponse of states with low NAEP response rates.

I. Background

Before we explore the methodology for estimating nonresponse bias and other related issues, it is useful to understand the NAEP methodology for estimation.

The basic survey design for a given state involved the selection of approximately 100 schools with probability proportionate to size of 8th grade enrollment. The schools were (implicitly) stratified based on urbanicity, percent of minority enrollment, and household income. The amount of stratification depended on the size of the state school sampling frame relative to the sample size. Larger states permitted greater implicit stratification. Some schools were selected with certainty.

If possible, given a school refusal to cooperate, a substitute school with similar characteristics was selected and assigned the probability of the originally sampled school. About 30 students were selected per sampled school. Students were excluded from the test-taking if they were incapable of taking the assessment. Exclusion rates for states ranged from 2 to 8 percent.

The school estimation weight was based on two factors: the inverse of the probability of selection and the school nonresponse adjustment factor. In general, when schools within a state did not respond and were not substituted for, nonresponse classes were created based on urbanicity, percent minority, and median household income (the same variables that were used for implicit stratification). Nonresponse adjustment classes varied from state to state depending on the distribution of these characteristics, absolute sample size within a potential nonresponse adjustment cell, and the size of the nonresponse adjustment factor. If any nonresponse class had fewer than six schools or a ratio greater than or equal to 1.35, it was collapsed (until the criteria were met).

A second set of weights was determined for each sampled school to be used for variance calculation. These weights were based on the jackknife replication procedure. In the jackknife procedure the impact of nonresponse and nonresponse adjustment are reflected in the variance estimate.

II. Simulations Based on State Assessment Tests

Because of the timing, the contractor, Synectics, focused on the 1990 states, contacting a number of them with school-level nonresponse. Of these states, the contractor found only two that conducted state assessments on all their schools and that were comparable to NAEP. The states were California and Illinois.

A. California

California provided a list of 1577 schools with addresses and school scores for its state assessment. This file was matched to the NAEP school sample of 104 schools. There were 6 nonrespondent schools and no substitute schools for the California NAEP assessment. The weighted correlation coefficient between the NAEP scores and the California Achievement Test (CAT) school scores was .92, based on the 98 responding schools. This indicates a fair amount of comparability, implying that results of the simulations based on the CAT scores would have similar application to the NAEP scores.

A number of possible estimates of bias were possible:

$\text{bias}_{1u} = M_{1u} - M^*$ and
$\text{bias}_{1a} = M_{1a} - M^*$, where

M_{1u} is the estimate of the CAT score for the respondent schools based on the inverse of the probability of school selection unadjusted for nonresponse.

M_{1a} is the estimate of the CAT score for the respondent schools based on the weight adjustments for nonresponse (i.e. we use the inverse of the probability of selecting the school times the weight for the nonresponse adjustment cell), and

M^* is the true average CAT school score taken over all schools in California.

To be a good reflection of NAEP bias, the above school scores have been student weighted. The average score using just the school weights would probably be similar, but not identical.

We can test whether these estimates of bias are significantly different from 0, using the estimated variances of M_{1u} and M_{1a}. The estimates of variance are based on the jackknife, the method used to calculate variances for NAEP. For California, the estimates were as follows:

$M_{1u} = 270.10$, $S_{1u} = 4.50$,
$M_{1a} = 269.65$, $S_{1a} = 2.96$,
$M^* = 272.94$,

$\text{bias}_{1u} = -2.84$, and
$\text{bias}_{1a} = -3.29$.

In this instance the nonresponse adjusted estimate of bias is larger than the unadjusted estimate. Based on the sampling variance, neither estimate of bias is significantly different from 0.

Another estimate is of interest is the unbiased estimate M_o (of the CAT score) based on the original (NAEP) sample of schools. This estimate for California, 269.43, is consistent with the thought that sampling error dominates school nonresponse bias.

B. Illinois

The NAEP sample size for Illinois was 105 schools, of which 23 schools were nonrespondents. Of these, 19 schools were ultimately substituted for, and 4 remained nonrespondents. Illinois tests eighth grade mathematics as a part of its Illinois Goals Assessment Program (IGAP). The weighted correlation coefficient between the NAEP scores and the IGAP state school scores was .93, based on the 101 responding schools. For Illinois we also estimated the unweighted standard deviations for the NAEP scores and the IGAP scores.

Because substitution was employed in Illinois, it is possible to calculate four IGAP estimates in addition to the universe estimate M*. The first two are the same as before — M_{1u}, based on the original sample cases and NAEP base weights, and M_{1a}, based on the initial respondents (excluding substitutes) and the appropriate NAEP nonresponse adjustments. The third is M_{2u}, based on the original respondents plus substitutes and the fourth is M_{2a} based on the original respondents plus substitutes with the weights adjusted for nonresponse.

With the new estimate M_{2a} we are able to separate out the "nonresponse" bias, if any, into two components — that due to nonresponse and using the NAEP nonresponse weight adjustment methodology, and that due to nonresponse plus substitution and using the NAEP nonresponse weight adjustments. The difference between the two could be considered an estimate of the impact of substitution.

The two new estimates give rise to two new estimates of bias:

$\text{bias}_{2u} = M_{2u} - M^*$ and
$\text{bias}_{2a} = M_{2a} - M^*$.

As before, these can be tested to see whether they are different from 0.

The results for Illinois are as follows:

$M_{1u} = 243.75$, $S_{1u} = 6.01$,
$M_{1a} = 248.59$, $S_{1a} = 4.79$,

$M_{2u} = 245.35$, $S_{2u} = 5.87$,
$M_{2a} = 244.88$, $S_{2a} = 4.19$,
$M^* = 248$,

$\text{bias}_{1u} = -4.25$,
$\text{bias}_{1a} = .59$,

$\text{bias}_{2u} = -2.65$, and
$\text{bias}_{2a} = -3.12$.

We can see that the estimate of bias with substitutes and nonresponse adjustments is greater in absolute value than the estimate based on no substitution but with nonresponse adjustments.

Given the size of the biases and their

estimated standard errors, we cannot reject the hypothesis that the biases are equal to 0.

The unbiased estimate M_0 (of the IGAP score) based on the original (NAEP) sample of schools for Illinois is 245.31.

C. Conclusion

While the above are only estimates of bias and ones that are not statistically significant given the small sample sizes, further research on the effects of school nonresponse at the state level would be useful, especially with respect to states with much higher rates of nonresponse than California or Illinois. In addition, the estimate for Illinois prior to substitution (but adjusting for nonresponse) has a smaller estimated bias than after substitution indicating that it might be useful to explore further the use of substitution.

III. Simulation Study of Nonresponse Using States with 100% School Response

Based on the 1990 NAEP state response rates it was decided to simulate nonresponse at three levels: 5%, 10%, and 20%. Three states with 100% response were used as the simulation states: Georgia, Colorado, and Connecticut. Because the school response rate was 100%, the published estimates for these states are unbiased estimates (unbiased at least, due to school nonresponse).

In order to simulate nonresponse, it would be necessary to take a nonrandom sample of schools, eliminate them from the file, and make estimates based on the NAEP estimation system. It would, therefore, not be sufficient to eliminate a random sample of schools in the state, nor would it be sufficient to eliminate a random sample of schools within a state within cells based on the nonresponse variables because both of these would result in unbiased estimates that would converge to the sample estimate as the number of simulations is increased.

As mentioned in the Background, the nonresponse adjustment cells were based on median income, percent minority, and urbanicity. Therefore, any nonresponse bias, if it exists, must be the result of some other variable being correlated with nonresponse, or, looked at another way, there must be a variable within the median income by percent minority by urbanicity nonresponse adjustment cells that exhibits one level for the respondents and another for the nonrespondents.

To put the results in context, the following are the estimates and standard errors from the 1990 NAEP Trial State Assessment of the NAEP average scores for Georgia, Colorado, and Connecticut:

State	Estimate	Standard Error
Georgia	258	1.3
Colorado	267	1.0
Connecticut	270	1.1

As mentioned earlier, this jackknife variance includes the variation due to nonresponse and nonresponse adjustment.

Initially, schools were eliminated from the states completely at random. Schools were formed into nonresponse adjustment classes and any necessary collapsing was performed. Then nonresponse adjustment factors were applied to the responding schools and the average score was calculated. While this method did not tell us anything about the bias, it did provide information about variation associated with nonresponse and subsequent nonresponse adjustment.

Table 1: Georgia simulation with 3 rates of nonresponse and 30 observations.

		Nonresponse	
Obs.	5%	10%	20%
1	258.300	258.201	257.502
2	258.648	258.218	259.865
3	258.082	258.333	257.578
4	258.334	258.769	258.747
5	258.498	258.505	258.566
6	258.531	258.075	257.470
7	258.702	259.116	257.733
8	258.545	257.612	258.874
9	258.301	257.938	259.621
10	258.548	258.253	258.529
11	258.246	258.832	258.514
12	258.787	257.465	258.034
13	257.865	258.670	258.771
14	258.646	257.965	257.890
15	258.184	258.337	258.043
16	258.137	257.821	257.660
17	258.301	258.614	256.776
18	258.603	258.372	258.445
19	258.576	258.555	258.067
20	258.251	257.521	258.243
21	258.446	258.163	256.290
22	257.949	257.571	259.379
23	258.569	258.580	258.533
24	257.609	256.885	258.024
25	258.268	257.936	259.216
26	258.629	258.105	258.754
27	258.105	258.214	259.214
28	257.898	257.605	258.847
29	258.630	258.286	257.001
30	258.314	257.977	257.119

Table 2: Bias and standard deviation with simulated nonresponse for Georgia and Colorado.

	Georgia		Colorado	
	bias	s.d.	bias	s.d.
5% Nonresponse	.1088	.2842	.0637	.1692
10% Nonresponse	.1014	.4693	.0699	.3160
20% Nonresponse	.0012	.8489	.0935	.4902

We calculated 30 simulations for each of the states at 5%, 10%, and 20% nonresponse rates (see Table 1 — for the state of Georgia).

Note: For readers familiar with NAEP, we mention that these results are for the first "plausible value," but similar results are available for the other four plausible values.

Observe that the simulated scores are quite close to the true value 258. (In fact, if we did enough simulations, the average of the composite scores would exactly equal the true value.)

In Table 2, for each simulation score, the true score has been subtracted, the difference squared, summed across the 30 simulations, and divided by 30. The first thing to note is that the variance increases as the percent nonresponse increases from 5% to 20%. This is what we would expect, because there is greater variability with more nonresponse.

The second piece of information that is somewhat instructive is the size of the variance (1.69) relative to the calculated standard deviation for Georgia of 1.3 (based on 100% response rate and no nonresponse adjustment). With nonresponse, two factors increase the variance: first, the variance is increased due to the decreased sample size and second, it is increased by the effect of nonresponse adjustments on weight variability. Therefore, if we take the standard deviation, approximately .8, and square it to get .64, then we might infer that the variance of the estimate for Georgia would be equal to 1.69 + .64 = 2.33, if the component of variance due to nonresponse and adjustment is assumed to be additive. Because a 20% nonresponse rate would inflate the variance by a factor of 1/.8 = 1.25, the resulting variance would be approximately 2.11.

Now, if we divide the total projected variance 2.33 by 2.11, we get a factor of 1.10 — an estimate of the increase in variance due to the application of nonresponse adjustment factors. This factor seems reasonable, and implies that the nonresponse adjustments only add about 10% to the estimated variances. (This factor could be verified another way by calculating the jackknife variance of one of the simulated samples.)

IV. Future Work

With respect to simulations based on state assessment tests, further research is desirable to see if there are other states that have tests comparable to NAEP conducted on all schools.

With respect to simulating the impact of school nonresponse on states with 100% response, one could base the simulated patterns of nonresponse on the patterns actually observed in states with less than 100% response. In 1992 one of the participating states had a nonresponse rate of 38%. One could simulate the impact of various levels of nonresponse between 5% and 45% and try to develop a sense of when the nonresponse becomes so large as to have a significant impact on the estimated results.

Acknowledgments

The authors are grateful to many people for their work or advice on this project. We thank particularly Glenn Galfond and Valerie Smith of Price Waterhouse; Arthur Kirsch, Sameena Salvucci, Boris Freidlin, Ramal Moonesinghe, and Al Holt of Synectics; Keith Rust of Westat; and Bob Burton, Roger Herriot, and Sam Peng of NCES. Special thanks go to Ramal Moonesinghe for invaluable contributions during the latter stages of the project.

*This paper is intended to promote the exchange of ideas among researchers and policy makers. The views are those of the authors, and no official support by the U.S. Department of Education is intended or should be inferred.