

PARTICIPATION STANDARDS FOR 12TH GRADE NAEP

James R. Chromy, RTI International

November 15, 2005

Table of Contents

Executive Summary	1
1. A New Vision for 12 th Grade NAEP	3
2. Response Experience and Current Standards	3
3. Response Rate Definitions	6
4. Overall Population Coverage	9
5. Reasons for Nonresponse	10
6. Coordination With Related Education Surveys	12
7. How Low Response and Coverage Rates Affect the Quality of Data	12
8. Usual Compensation Measures	13
9. Item Nonresponse	14
10. Conclusions and Recommendations	15

Executive Summary

The National Assessment Governing Board's new vision for grade 12 NAEP poses serious challenges for dramatically improving response rates at the school and student level as a means to guarantee the quality of NAEP grade 12 estimates.

Declining survey response is a problem in voluntary surveys in general. Recent NAEP response rate experience verifies that response rates without mandatory public school participation continue to decrease reaching a low 55 percent overall weighted response rate for all grade 12 students in 2002. In addition, the overall response rate for students attending nonpublic schools dropped to below 50 percent in 2002. Clearly, some decision must be reached about the advisability of continuing the grade 12 assessment in the presence of these low response rates.

Existing standards and guidelines provide some guidance. Note that grade 12 response rates are near the levels which trigger serious review of the viability of the program under current NCES standards. Since even the NAEP guidelines are not satisfied by the national grade 12 assessment and prescribe only annotation of the doubtful quality of the resulting survey estimates, they are only useful in identifying a potential problem without providing specific details about its resolution. The decision to continue should be based on the development and testing of survey protocols which are expected to increase the response rate to more nearly acceptable levels in future NAEP survey rounds. It should also be based on an analysis of nonresponse utilizing additional auxiliary data that either (1) reassures us that the probable bias due to nonresponse is not serious enough to curtail future assessments at grade 12 or (2) identifies and defines the actual impact of low response rates on NAEP estimates. A decision to curtail all or part of future grade 12 assessments might include temporarily redefining the grade 12 target population to cover only students attending public schools while an effective strategy is being developed to survey the nonpublic segment.

Methods to increase response rates should be guided by what we already know about the reasons for nonresponse. In some cases, these data are already being collected by the field staff. Available data need to be organized and analyzed; based on this preliminary analysis, recommendations should be developed on how best to increase response or how to obtain additional information to better understand the response process and the reasons for nonresponse. Experimental studies should be developed to test the effectiveness of proposed methods. Some of these experiments could be embedded in a future round of the grade 12 assessment.

An overall quality profile for NAEP should be developed and updated annually with focus on potential nonresponse bias. Some auxiliary data on responding and nonresponding schools and students already exist. The need for and potential availability of additional data should be seriously explored. Population coverage issues for the disabled (SD) and limited English proficient (LEP) students were investigated much more effectively after supplemental questionnaires completed by

school staff were used to obtain data on both participating and nonparticipating SD/LEP students. Similar methodologies could be applied to study student nonresponse by collecting available auxiliary data for both respondents and nonrespondents.

Issues of item response should also be part of any future research agenda to improve the quality of NAEP estimates.

The overall response problem is serious and threatens the future viability of grade 12 assessment. It should be addressed by a coordinated and continuing research agenda that addresses nonresponse problems for all grades, but initially places special emphasis on grade 12. Given the resources required to conduct NAEP in its present form, the provision of additional resources, or the allocation of a portion of the currently available resources, to address the response rate problem in an organized and scientific manner is fully justified and, in fact, essential to the successful continuation of the NAEP statistical data series.

PARTICIPATION STANDARDS FOR 12TH GRADE NAEP

James R. Chromy, RTI International

1. A New Vision for 12th Grade NAEP

In its report to the National Assessment Governing Board (NAGB), the National Commission on NAEP 12th Grade Assessment and Reporting (March 5, 2004) proposed a new vision for 12th grade NAEP and recommended changes to increase its relevance and usefulness to policymakers and the public. With regard to participation, they recommended that:

“NAEP’s leaders—the National Assessment Governing Board, the U.S. Department of Education, and the National Center for Educational Statistics—should develop and implement bold and dramatically new incentives to increase the participation of high schools and 12th grade students in NAEP and the motivation of 12th grade students to do their best on NAEP.”

The commission also recommended expanding state assessment to include 12th grade in reading and mathematics and to make participation mandatory for all states.

2. Response Experience and Current Standards

Much like the experience of other surveys, the NAEP combined school and student response rates have been falling in the last few years. Tables 1a and 1b show recent experience for all three grades¹. In 2003 and 2005, the school response rates for grade 4 and 8 public schools were at or near 100 percent in most states as a result of the No Child Left Behind Act requirements for mandatory participation.

Table 1a. Combined School and Student National Response Rates Before Substitution by Grade: NAEP Reading/R3, 1998 to 2005

	1998	2002	2003	2005
Grade 4	78%	79%	92%	90%
Grade 8	71%	75%	89%	88%
Grade 12	56%	55%	NA	55%

Source: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1998, 2002, 2003 and 2005 Assessments.

¹ Tables 1a and 1b are an update of an earlier draft version of Table 1; the updated results provide response data through 2005 and were provided by Young Chun, ESSI, and Andrew Kolstad, NCES.

Table 1b. Combined School and Student Response Rates of the National Public Before Substitution by Grade: NAEP Mathematics/R3, 1996 to 2005

	1996	2000	2003	2005
Grade 4	81%	82%	94%	93%
Grade 8	75%	76%	91%	90%
Grade 12	62%	60%	NA	57%

Source: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 1996, 2000, 2003 and 2005 Assessments.

Unless grade 12 participation is made mandatory, the lower response rates at grade 12 are likely to continue. To get a better understanding of the combined response rate it is useful to examine the school and student components of response separately. Table 2 shows selected results from 2002 NAEP for grades 4, 8, and 12. Response rates in 2002 decreased with increasing grade at both the school and student levels. School response was a serious problem at all grade levels, but the problem with student response rate was unique to grade 12. This is the likely explanation for the historically lower overall response rates shown for grade 12 in Tables 1a and 1b.

Table 2. Weighted Response Rates: 2002 NAEP Reading

	Grade 4	Grade 8	Grade 12
Weighted school response rate before substitution	84%	82%	74%
Weighted student response rate after make-up sessions	94%	92%	74%
Combined school and student response rate	79%	75%	55%

Source: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, The Nations Report Card: Reading 2002, p. 138.

Current NCES standards (Standard 2-2-2) require that “unit response rates must be calculated without substitution and that survey data collections be designed to meet unit level response rate parameters that are at least consistent with historical response rates from surveys conducted with best practices.” Guideline 2-2-2A provides current NCES historical experience. For assessments, this guideline specifies 80 percent response rates at the school level and 85 percent at the student level. Note that the 2002 response rates shown in Table 2 satisfy the guidelines for both school and student response rates for grades 4 and 8, but not for grade 12. NCES standard 2.2.4 states that a nonresponse bias analysis is required at any stage of data collection with a response rate less than 85 percent. Based on this standard, a nonresponse bias analysis would be required at the school level at all three grade levels and at the student level at grade 12. Standard 2-2-5 further specifies that “where prior experience suggests the potential for an overall unit response of less than 50 percent, the decision to proceed with data collection must be made in consultation with the Associate Commissioner, Chief Statistician, and Commissioner” (NCES Statistical Standards, 2002). The overall grade 12 response rate is most disturbing; if trends continue, it could be difficult to justify conducting the grade 12 assessment at all.

A better understanding of the school and student response can be obtained by looking at response rates for public and nonpublic schools. Even without mandatory participation at grades 4 and 8 in NAEP 2002, school response rates tended to be higher for public schools as shown in Table 3. In contrast to the results for schools, Table 3 shows that students tend to respond at higher rates in nonpublic schools. A better understanding of the reasons for nonresponse in each circumstance is needed before any general approach can be identified to increase overall response rates.

Table 3. School and Student Response Rates by School Type: NAEP 2002¹

Grade	Public Schools		Nonpublic Schools		Combined	
	School Response Rate	Student Response rate	School Response Rate	Student Response rate	School Response Rate	Student Response rate
4	85%	94%	74%	95%	84%	94%
8	83%	91%	68%	95%	82%	92%
12	76%	72%	55%	88%	74%	74%

Source: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, The Nations Report Card: Reading 2002, p. 138.

¹ School response rates are computed before substitution.

Additional guidelines (See Appendix A, NAEP 2002 Report Card, p. 172 ff.) were developed for reporting by jurisdiction (state or territory) results for the public school portion of the sample. Guideline 1 required a weighted school participation rate of at least 70 percent. If Guideline 1 was not satisfied by a jurisdiction, NCES did not prepare a separate report for that jurisdiction. NAGB specified that in such cases, results should be reported “in a different format”. In addition, the Education Information Advisory Committee (EIAC) passed a resolution calling for data from such jurisdictions not to be published. In 2002, two states did not meet guideline 1 at grade 4, and four states did not meet it at grade 8.

Guideline 2 pertained to jurisdictions that satisfied guideline 1. It specified that jurisdiction results would receive a special notation if the weighted public school participation rate was less than 80 percent before substitution and less than 90 percent after substitution. In 2002, nine states did not meet guideline 2 at grade 4; eight states did not meet guideline 2 at grade 8.

Guideline 4 applied to jurisdictions meeting guideline 1. It specified that jurisdiction results would receive a special notation if the weighted student response rates were below 85 percent.

Guidelines 3 applied to jurisdictions meeting guideline 1 and specified participation rates at the school level for important classes of schools. Classes were defined by degree of urbanization, minority enrollment, and median household income of the school area. Within each class, an 80 percent weighted school response rate (after substitution) was required if the nonparticipating schools in the class accounted for more than 5 percent of the jurisdiction’s total weighted sample of public schools.

Guideline 5 applied to jurisdictions meeting guideline 4 and specified participation rates at the student level for important classes of students. Classes were defined by age, disability status, LEP status, type of assessment session, school level of urbanization, school minority enrollment, and median household income of the school area. Within each class, an 80 percent weighted student response was required if the weighted count of nonresponding students in the class accounted for than 5 percent of the jurisdiction's total weighted sample of public school students.

Failure to meet guideline 3 or 5 required that the jurisdiction's result be annotated. These guidelines were implemented to protect against bias from pockets of nonresponse in particularly important classes of schools or students.

It should be noted that these guidelines pertain only to state (or jurisdiction) assessments that are conducted for grades 4 and 8 and only to public schools. It should also be noted that if the guidelines were applied to grade 12 national results for public schools, they would satisfy guideline 1, but not guideline 2 since the school response rate after substitution was approximately the same as before substitution. The weighted student participation rates for public school students also would not satisfy guideline 4. If the jurisdictional guidelines were applied to the national grade 12 sample of nonpublic schools, they would fail guidelines 1 and 2 for school participation, but would pass guideline 4 for student participation. Guidelines 3 and 5 cannot be easily evaluated in either case. Were the guideline to actually apply to grade 12 national results for public schools, the separate publication of results would be in doubt (Guideline 1). If they were published, they would require annotation to indicate the possibility of serious bias resulting from low response rates at both the school and student levels (Guidelines 2 and 4). While the student response rate for the nonpublic schools would be acceptable the overall response considering both school and student response rates would be below 50 percent. If this low level of overall response rate were expected to continue in future assessments, the decision to proceed with data collection would need to be made in consultation with the Associate Commissioner, Chief Statistician, and Commissioner" in order to comply with NCES Statistical Standard 2-2-2.

This discussion of grade 12 response experience and its relation to existing guidelines points out the seriousness of the response rate problem.

The remainder of this paper discusses some technical issues related to survey response and a set of recommendations for dealing with NAEP nonresponse.

3. Response Rate Definitions

In the simplest case, a response rate measures the portion of the selected sample that responds by completing a questionnaire. This simple definition works well when a single stage sample is used and all sample members are selected with equal probability. NAEP requires multi-stage sampling and higher sampling rates (oversampling) for special populations such as minority students, disabled (SD), limited English proficient

(LEP) students, and students attending nonpublic schools. The higher sampling rates result in larger sample sizes for these groups permitting more precise estimates that pertain to them specifically; they also result in an overall design with a variety of planned sampling rates. Even without the planned variation in sampling rates, some exceptions to the planned rates occur due to the imperfection of advance information about the size of selected schools and the need to adjust the sampling rates during the final stages of sample selection to limit burden on individual schools or for other reasons.

Since the purpose of sampling is to obtain estimates about a population rather than just about the sample, anyone evaluating the quality of survey data based on response rates should be concerned about the proportion of the population represented by the responding sample rather than simply the proportion of selected sample. For this reason, organizations such as the American Association for Public Opinion Research (AAPOR) and the Council of American Survey Research Organizations (CASRO) recommend weighted response rates for complex surveys. NCES standards (Standard 1-3) specifically require the use of weighted response rates for evaluating and comparing NCES surveys; unweighted rates may be used for monitoring field operations. Weighted response rates are and should continue to be the ultimate basis for evaluating NAEP data quality based on response.

In multi-stage surveys, it may not be possible to determine the number of subjects selected at the final stage when nonresponse occurs at an earlier stage. When schools do not respond, it is necessary to estimate the impact of the school response rate on the overall response rate for students or other subjects sampled within schools. This is typically done by computing the response rate at each stage of sampling and computing the overall response rate as the product of these two rates. For the NAEP sample this requires taking the product of the school response rate and the student response rate to obtain an overall response rate.

Using weights in computing response rates can increase or decrease the response rates relative to unweighted results. As an example, the NAEP sample design requires sampling of SD/LEP students. Even after allowing for exclusions, the response rate for these students is generally lower than that of other students. Weighted response rates insure that their lower response rates contributes to the overall response rate only to the extent that SD/LEP students are part of the overall eligible population. Weighted response rates also eliminate the opportunity for “gaming” the sample allocation process to produce high average response rates by concentrating the sample in subgroups with known higher response rates.

The NAEP sample design selects one or two closely matched substitute schools for each sample school (1998 Technical Report, Chapter 3) when feasible. When a school refuses to participate, attempts may be made to recruit one of the pre-selected substitutes to replace the refusing school. If the substitute agrees to participate, the school is treated as a respondent, but the denominator of the response rate remains unchanged. The following example is taken from the NAEP State Assessment Sample Design FAQ on the NAEP web site.

Example:

“Consider the following example:

- 105 schools selected in original sample;
- 5 close, or don’t have relevant grade;
- 87 original sample schools participate; and
- 13 original sample schools refuse.

For the 13 refusals, there is no substitute in the system for 2. For the 11 refusals that do have substitutes:

- 6 substitutes participate;
- 2 refuse, and
- 3 are never contacted.

$$\text{Before substitution response rate} = \frac{87}{(105 - 5)} = 87\%.$$

$$\text{After substitution response rate} = \frac{(87 + 6)}{(105 - 5)} = 93\%.$$

(These figures are only approximate because weighted response may differ somewhat from these numbers.)”

By this device, substitution can only increase the response rate. We could just as well treat the 11 cases available for substitution as an additional sample release. Then for these 11 we have 3 “no contacts” (a form of nonresponse), 2 refusals, and 6 respondents. By treating all 11 as an additional sample release, the after substitution response rate would be $\frac{(87 + 6)}{(105 + 11 - 5)} = 84\%$. It is not clear why the 3 schools not contacted were not contacted and it is possible that some of them might be ineligible (closed or no relevant grade). But this alternative produces a lower, not higher, response rate after substitution. Empirically, the chances of getting substitute schools to participate can be expected to be lower than the chance for initially selected schools. The initial sample school nonresponse decision may occur fairly late in the process allowing inadequate time to make contacts and seek cooperation from substitute schools.

For evaluation purposes, NAEP follows NCES standards and reports the before substitution weighted response rate. The use of pre-selected augmented subsamples that can be applied in design strata with low sample yield is not uncommon. These types of sample supplements are often used in other surveys to compensate for unexpected low eligibility as well as for nonresponse and are viewed as a method of insuring an adequate sample size. When used in this manner, the entire additional sample release contributes to the denominator of the response rate and the respondent proportion contributes to the

numerator of the response rate. The method is not necessarily expected to increase the response rate, but it does allow for matching sample shortages with new sample from the same sampling stratum. If substitution or sample augmentation is to be used, then a more appropriate response rate would properly reflect the total added sample in the denominator of the response rate calculation.

At the student level, make up sessions are scheduled when the response rate falls below a specified level. In 1998, this level was 90 percent. This is not a sample augmentation method. It is going back to initial nonrespondents who were absent for the scheduled assessment date and time and giving them an opportunity to participate at a later time. In this case, the additional response adds directly to the number of respondents and does not increase the selected sample size. As a result, a legitimate increase in response rate occurs.

4. Overall Population Coverage

This report focuses on participation or response rates which relate the responding sample to the selected sample. The sample is selected from a frame and the frame is designed to match the study population. The study population for NAEP includes fourth-grade, eight-grade, and twelfth-grade students in public and nonpublic elementary and secondary schools. The survey is conducted during the period January to March. Student eligibility presumably depends on their being enrolled in an eligible school at the time the survey is conducted. Students who are known to have dropped out of school are not eligible. At grade 12, there may be an unknown number of students who have not officially dropped out, but are less than fully engaged in full time attendance; these types of students are the ones most likely to be absent on the day of NAEP administration. Schools with year-round programs may have a portion of their grade 12 enrollment in a break period at the time of assessment (This applies to other grades also). Other twelfth-grade students may be attending school on a part time basis to finish minimum requirements for graduation while they work or pursue college courses. It is likely that the frame includes some students that are not really in the study population and unnecessarily penalizes the response rates as calculated. It is also possible that some population eligible students are excluded from the frame due to the timing of school breaks and NAEP data collection.

School samples are drawn well in advance of the testing period to allow time for pre-survey contacts and scheduling activities. As a result, some new schools may not be included in the list of schools serving as a frame. Other schools in the frame may no longer be eligible. The NAEP sampling and field procedures include processes to add new schools and allow for selected schools to be ineligible.

Exclusion rules for SD and LEP students have been updated to be more inclusive and to provide accommodations in many cases. This expansion of the target population has added a component that has had lower response rates and this may be contributing to some, clearly not all, of the decrease in response rates over time. Public schools tend to

have a higher proportion of SD/LEP students and this may explain some (not all) of the difference in student response between public and nonpublic schools.

Statistical inference under ideal conditions allows the results from a sample to be extended to the frame from which the sample was selected. It is generally assumed that the frame (list of schools and lists of students within schools) corresponds very closely to the conceptual target population. This correspondence is not automatic and should not be assumed. A new vision for 12th grade NAEP should include a careful evaluation of the grade 12 target population definition in the context of current educational practices and patterns of school attendance during the assessment period.

Just as overall response rates are computed as the product of school and student response weights, an overall population coverage can be viewed as the product of the overall response rate and the frame coverage rate.

5. Reasons for Nonresponse

If response rates are to be increased in the future, more needs to be known about the reasons for nonresponse. Some information is already known as illustrated in Tables 4 and 5 below. Table 4 shows some reasons for nonresponse at the school level. Note that the “out of scope” category does not count against the response rate. District refusal for all schools in the district or for a particular school appears to be a serious problem. More needs to be known about the reasons that districts refuse. Is it because of too much testing already in their school schedules, general resistance to federal programs, or other reasons? From this table it is not clear whether the individual school refusals are mostly nonpublic schools, but this appears quite likely. More detailed information is needed to help develop targeted strategies. This type of information is currently being collected from field staff for refusing private schools and from state coordinators for refusing public schools and districts. This information needs to be organized, analyzed, and reported. General impressions from examining the data are that “too much testing” is a major reason for school nonresponse in general and that, for private schools, “lack of useful information” also ranks high².

	Grade 4	Grade 8	Grade 12
Original sample	441	465	522
Out of scope	24	46	60
District refusal for all schools	18	19	25
District refusal for this particular school	11	14	17
School refusal	47	43	75
Participating	341	343	345

Source: 2001 NAEP, national main assessment.

²Communication from Nancy Caldwell (Westat E-mail of October 13).

Table 5 shows two general reasons for student nonresponse: absent and refused. The reasons are shown by school characteristics. The startling result is the overall differences in student response between public and nonpublic schools and the large impact of absence on the public school student nonresponse. In order to develop procedures to improve student response, more information is needed about the underlying reasons for nonresponse within these broad categories. Are absent students away from school or involved in other school activities at the scheduled time? Are refusals based on parental refusals or individual student refusals? Was a written parental consent required by the school administration? Was active or passive consent required? Currently, data on reasons for absences are recorded on the administration schedule; field staff record reasons by code for temporary absence, long-term absence, chronic truant, disruptive behavior, parental refusal, student refusal, and “other specify”³.

Table 5. Reasons for Student Nonresponse by School Characteristic:2000 NAEP Grade 12 Mathematics and Science Assessments

	Percent Assessed	Percent Absent	Percent Refused
School type			
Public	75.4	20.2	4.4
Private	93.8	5.5	0.7
School Location			
Urban	68.4	23.3	8.3
Suburban	76.4	20.1	3.5
Rural	87.5	11.6	0.9
Census Region of School			
Northeast	76.2	20.0	3.8
Midwest	76.0	22.3	1.7
South	79.3	15.4	5.3
West	68.9	24.7	6.5
Percent Students in NSLP			
0-10%	77.5	18.2	4.4
11-50%	78.6	18.5	2.9
51% or Higher	77.7	16.2	6.1
12th Grade Enrollment			
0-100	90.4	8.9	2.7
101-500	75.3	19.6	5.1
501 Or More	77.3	19.6	3.1
Percent Repeating 12th Grade			
0	81.0	16.6	2.4
1 to 2	77.6	18.1	4.3
3 or higher	75.2	19.0	5.9

Source: Young Chun (2005), "A Theory-driven Approach to Explaining and Reducing Twelfth Grade Non-participation in NAEP," paper presented at the annual conference of the American Educational Research Association, Montreal Canada.

None of the more detailed school or student reasons for nonresponse data have been examined in the preparation of this report, but such an examination and continued

³ Communication from Nancy Caldwell (Westat E-mail of October 13).

tracking of this data over repeated assessments will be crucial in developing effective strategies to increase response rates at both the school and student levels.

6. Coordination With Related Education Surveys

Other educational surveys share the general problem of decreasing response rates with NAEP. A major factor in limiting school participation is an already full schedule of testing programs required at the state level by NCLB or other initiatives. The survey program that gets admitted into a school is often the one that asks first. With a long continuing planning horizon, NAEP is in a strong position to be first among the non-mandatory programs.

Other NCES-sponsored studies have, on occasion, coordinated school sample selection with NAEP so that sample overlap and individual school burden could be minimized. Such coordination is not effective when the sampling rate in either survey is very high. As an example, some state NAEP surveys involve all or almost all schools in the smaller states.

Coordination of survey samples and annual scheduling of Federal surveys to avoid several programs in the field simultaneously could help improve school response rates for all the surveys involved.

7. How Low Response and Coverage Rates Affect the Quality of Data

It is important to understand how poor response rates affect the quality of estimates. The damage is generally expressed in terms of the bias. The bias due to nonresponse can be written as $B_{nr} = (1 - r)(\mu_r - \mu_{nr})$, where r is the response rate, μ_r is the expected value of one of the survey estimates for respondents, and μ_{nr} is the expected value of the same survey estimate for nonrespondents. It is difficult to know the expected value of the estimate for nonrespondents since we have no data about them. If the response rate is kept high, we can be sure that the bias will be low.

Since NAEP produces a large number of survey estimates for each sample, the bias may be different for each estimate. NCES standards prescribe nonresponse bias analysis when response rates do not meet guideline levels as discussed above. One way to conduct such an analysis is to identify certain characteristics or measures that are known for both respondents and nonrespondents. Often such measures come from the sampling frame from which the sample is selected. For these known measures, it is possible to compute the full sample estimate, say \bar{x} and an estimate based on respondents only. The estimate based on respondents only can be constructed two ways: (1) based on the original design-based weights (designated by $\bar{x}_{r,db}$), and (2) based on the final analysis weights after adjustments have been made to attempt to compensate for the nonresponse (designated by $\bar{x}_{r,adj}$). Two estimates of nonresponse bias can then be calculated as $\hat{B}_{nr,db} = \bar{x}_{r,db} - \bar{x}$ and $\hat{B}_{nr,adj} = \bar{x}_{r,adj} - \bar{x}$. The first shows the potential for

nonresponse bias with no weight adjustment; the second provides an estimate of nonresponse bias after weights have been adjusted. Keith Rust provided estimates of the second type for nonresponse at the school using known school-level variables and at the student levels using known student variables⁴. With a few exceptions, the estimated biases were not significantly different from 0. Given the large number of statistical tests conducted, his results are comforting. Theoretically, it might be possible to develop weight calibration methods that reduce the calculated bias to zero or near zero for all known variables. Even if this were done, it might not capture the dimension that determines student performance on NAEP. Any calibration method that does not capture the expected NAEP performance may do more damage to survey estimates by increasing the sampling error as a result of unequal weighting.

To examine the potential effect on NAEP measures that may or may not be associated with the control variables discussed above, Rust simulated purely theoretical measures of bias on the mathematics scale (standard deviation = 35 scale points) as a function of the school and student response rates for fixed values of the intraclass correlation coefficient and two correlation coefficients⁵. The first correlation coefficient measures the correlation between school response and school mean achievement; the second measured the correlation between student response and student achievement. Some general results include the following. For both response rates at 100 percent, there is no bias. For a school response to achievement correlation of 0, no bias is introduced by school nonresponse. For a student response to achievement correlation of 0, no bias is introduced by student nonresponse. Correlations of achievement with response rates have the most effect when both are in the same direction (both positive or both negative). For both school and student response rates at 85 percent (NCES guideline minimums), the simulated biases range from 0.8 scale points to 4.2 scale score points. If we allow the overall response rate to drop to 49 percent, say with school and student response rates both at 70 percent, the same range of simulated biases becomes 0.5 to 7.6 scale points.

Any scale point bias greater than 1.0 is disturbing because the magnitude of the bias is likely to exceed the sampling error and has the potential to seriously jeopardize the validity of comparative analyses. The problem is more serious when the groups being compared are subject to different school and student response rates. This type of situation has been noted for students attending public vs. nonpublic schools.

8. Usual Compensation Measures

Nonresponse adjustments and poststratification adjustments are often applied in an attempt to minimize nonresponse and overall coverage bias. Both methods are used for adjusting NAEP weights. The nonresponse adjustments insure that respondent weight

⁴ Keith Rust (2004), An Analysis of the Impact of School and Student Nonresponse on Sample Representation for the 2002 Grade 12 NAEP. Westat, May 4, 2004 presentation materials.

⁵ Keith Rust (2005), Potential Nonresponse Bias in NAEP Results. Westat, May 4, 2005 presentation materials.

totals for selected subgroups correspond to similarly defined totals for the selected sample. Nonresponse adjustments are based on variables known for all selected schools or for all selected students. Poststratification uses external totals based on larger samples (*e.g.*, the Current Population Survey) to force correspondence of weight sums for selected groups with external known totals. Variables used for poststratification only need to be known for the respondent sample, but population totals based on those variables must be known from the external sources. Poststratification has the potential to adjust for frame coverage as well as nonresponse. It is important to use external data that are much more precise than the NAEP estimates and that are based on the same eligible population definition.

Response bias analyses are conducted to provide some comfort about the likely level of nonresponse and also to guide the selection of variables to use in the nonresponse adjustment process. More relevant response bias analysis studies can be conducted if the known variables for respondents and nonrespondents relate more closely to student achievement.

In dealing with the coverage issue for SD/LEP students, auxiliary data were first selected in an excluded student questionnaire that provided information on the noncovered population, but could not be compared to the covered group. Subsequently, similar auxiliary data were collected for all initially identified SD/LEP students. The auxiliary information on both covered and not covered SD/LEP students could then be compared and used to study potential bias due to exclusion. In terms of the potential nonresponse bias study by Rust discussed above, we suspect that response is also positively correlated with achievement, at least for some student groups. Since the reasons for student nonresponse are different for different student groups, it may be possible to better utilize auxiliary information related to achievement. Such data would have to be collected for both respondents and nonrespondents. Rust indicates some suspicion that auxiliary information based on the transcript study is positively correlated with achievement data and could be used effectively to adjust for nonresponse.

9. Item Nonresponse

Overall survey response includes both unit response and item response. For NAEP, unit response is based on a combination of school and student response rates. A complete unit response is generally identified in terms of some minimum quantity of data obtained in a questionnaire or assessment instrument. Nonresponse weight adjustment procedures set the weights for nonrespondents to zero and adjust respondent weights to force them to sum to the level of the selected sample. For some unit respondents, additional nonresponse occurs at the item level. The assessment requires both background data and performance measures and the two types of data may be treated differently. Background data may be imputed if most of it is available. Assessment items are considered wrong if left blank when the respondent continues to attempt subsequent items. Blank items are treated as “not reached” and, therefore, missing when no subsequent items are attempted. The plausible scoring methodology implicitly

handles missing items the same way as items not in the package by design. A student's measured performance may be influenced by his/her views about improving test performance by answering few questions carefully or as many questions as possible by guessing if necessary. Of course, some students may cooperate to the extent of agreeing to participate, but then do not seriously apply themselves to completing the assessment.

The impact of item participation on the reliability of NAEP data is highly related to the scoring method and the plausible value methodology and needs further study by experts in those areas.

10. Conclusions and Recommendations

This discussion focuses on the following issues:

- Are additional standards needed for NAEP response rates?
- What can be done to improve response rates?
- What can be done to understand the resulting data quality and report to data users?
- What can be done to adjust for the remaining nonresponse?

Additional Standards: High quality data through the highest response rates that can be achieved should be the NAEP goal. The goal must be moderated by constraints on time and budget. Putting exact values or lower bounds on acceptable levels of nonresponse can lead to complacency once those levels are reached. Opportunities exist to control costs or to complete work on schedule by “letting up” on procedures designed to improve response rates once those minimum levels are reached. This concurs with the conclusion reached by a National Institute of Statistical Sciences Workshop⁶.

NCES standards and guidelines already exist. NAEP should be careful not to treat these guidelines as being adequate to insure high quality data. Unfortunately, response rates at grade 12 do not meet the standards and guidelines prescribed for jurisdiction level samples of public schools. No guidelines have been developed for treatment of data from nonpublic schools. Operationally, the NAEP goal should be to develop and implement effective procedures to solicit and achieve that high response. As long as response is voluntary, we should expect to have some nonresponse and be prepared to meet the challenge of keeping response rates at acceptable levels while respecting the rights of respondents to refuse participation.

At some low level of overall response, the wisdom of continuing a repeated survey program must be considered. NCES standards set this level at 50 percent and then

⁶ James Chromy, Graham Kalton, Roderick Little, Ingram Olkin, Lyle V. Jones, Valerie S. Williams, Johnny Blair, and Richard Jaeger, *Accounting for Missing Data in Education Surveys: A Workshop Report*, Technical Report # 19, National Institute of Statistical Sciences, Research Triangle Park, NC, February 1999.

require review by the Chief Statistician, Associate Commissioner, and Commissioner. Taken separately, the overall response for grade 12 nonpublic schools has already reached this threshold. Deliberations about continuing survey programs at this low level of response must focus on the resulting quality of the estimates produced using the resulting data. The most serious overall response rate is for the class of students attending nonpublic schools. A decision might be reached to forgo coverage of the nonpublic school component at grade 12 because it is not possible to obtain acceptable response rates to control biases at a reasonable level. This would still leave a problematic level of response for the class of grade 12 students attending public schools. This is basically using a target population re-definition to resolve the most serious part of the nonresponse problem.

Another opportunity to increase coverage of grade 12 would be to reschedule the grade 12 assessment to the fall semester of the senior year. This, of course, would measure students' achievement at a different point in their educational experience, but might be effective in increasing coverage of grade 12 students who may or may not complete their high school senior year. Changing the population definition in this way would require simultaneous samples at both test periods to calibrate trend estimates.

Improving Response Rates: In order to improve response rates, NAEP should first assemble and analyze the information needed to understand the reasons for nonresponse at both the school and student levels. As noted above, much of this information is already available from the field staff reports. If not available in a useable form, procedures should be implemented to collect, assemble, and analyze this information in future assessments. If the available data are not adequate or need further explanation, it may be necessary to go back to nonresponding public school superintendents and school principals to probe for the underlying reasons for school refusals to participate. By examining this type of information, it should be possible to develop a set of strategies to deal with particular subsets of the selected sample. As an example, it is clear that different strategies should be (and probably are) used for public vs. nonpublic schools. Catholic schools may require different approaches than independent or unaffiliated private schools. Large public school districts that reoccur in the sample on a regular basis clearly need special attention. While many of these procedures may already be in place, a major effort to improve response rates needs to review the current procedures and update them if necessary.

If overall survey or school testing burden is found to have significant impact on district and school cooperation, then some efforts to coordinate the external demands on schools from NCES surveys could benefit the response rates for all of the surveys involved. This has already been done to a limited extent with some of the surveys. A number of surveys involving schools are conducted by other government programs to study issues such as crime, youth smoking, health issues, and substance abuse. Other school-based surveys are conducted at the local and state government level as well as by universities and other organizations. Overall burden is likely to be a major issue.

Another issue that is likely to arise is the use of incentives (monetary or otherwise) to gain school support or compensate for additional required efforts. Incentives to students should also be investigated.

As new procedures to improve response rates are developed, experimental studies should be conducted to check their general feasibility and potential effectiveness.

Understanding and Reporting on Data Quality: An overall assessment of data quality for NAEP needs to include measures of error arising from nonsampling errors as well as the sampling error estimates normally provided. Nonresponse is recognized as a major contributor to nonsampling error. Overall survey quality profiles have been developed and published for many major government surveys. If this has not been done for NAEP, it should be done.

Nonresponse bias analysis would be part of the overall quality profile. The usefulness of a nonresponse bias analysis depends on the availability of auxiliary data for both respondents and nonrespondents. Serious consideration should be given to collecting, organizing, and analyzing such data for measures that are more likely to be directly related to NAEP student achievement measures.

Adjusting for the Remaining Nonresponse: Procedures are already in place to adjust for nonresponse. The opportunity exists to identify or collect additional auxiliary data at the student level to improve this process. The choice of variables to use in this way should be made based on a thorough nonresponse analysis involving a set of candidate auxiliary variables. Just as the procedures for improving response may need to be targeted for specific population subgroups, some of the procedures for nonresponse adjustment may also need to target population segments defined by the available auxiliary data.