Program for International Student Assessment (PISA) 2003:
U.S. Nonresponse Bias Analysis

September 2009

NCES 2009-088
U.S. DEPARTMENT OF EDUCATION

ies NATIONAL CENTER FOR EDUCATION STATISTICS
Institute of Education Sciences

Page intentionally left blank.

Program for International Student Assessment (PISA) 2003: U.S. Nonresponse Bias Analysis

September 2009

David Ferraro
Jennifer Kali
Trevor Williams
Westat

Edited by:
Lydia Malley
Education Statistics Services Institute – Child Trends

Eugene Owen
Project Officer
National Center for Education Statistics

NCES 2009-088
U.S. DEPARTMENT OF EDUCATION

U.S. Department of Education
Arne Duncan
Secretary

Institute of Education Sciences
John Q. Easton
Director

National Center for Education Statistics
Stuart Kerachsky
Acting Commissioner

The National Center for Education Statistics (NCES) is the primary federal entity for collecting, analyzing, and reporting data related to education in the United States and other nations. It fulfills a congressional mandate to collect, collate, analyze, and report full and complete statistics on the condition of education in the United States; conduct and publish reports and specialized analyses of the meaning and significance of such statistics; assist state and local education agencies in improving their statistical systems; and review and report on education activities in foreign countries.

NCES activities are designed to address high-priority education data needs; provide consistent, reliable, complete, and accurate indicators of education status and trends; and report timely, useful, and high-quality data to the U.S. Department of Education, the Congress, the states, and other education policymakers, practitioners, data users, and the general public. Unless specifically noted, all information contained herein is in the public domain.

We strive to make our products available in a variety of formats and in language that is appropriate to a variety of audiences. You, as our customer, are the best judge of our success in communicating information effectively. If you have any comments or suggestions about this or any other NCES product or report, we would like to hear from you. Please direct your comments to

National Center for Education Statistics
Institute of Education Sciences
U.S. Department of Education
1990 K Street NW
Washington, DC 20006-5651

September 2009

The NCES World Wide Web Home Page address is http://nces.ed.gov.
The NCES World Wide Web Electronic Catalog address is http://nces.ed.gov/pubsearch.

Suggested Citation
Ferraro, D., Kali, J., and Williams, T. (2009). *Program for International Student Assessment (PISA) 2003: U.S. Nonresponse Bias Analysis* (NCES 2009-088). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC.

For ordering information on this report, write to
U.S. Department of Education
ED Pubs
P.O. Box 1398
Jessup, MD 20794-1398

or call toll free 1-877-4ED-Pubs or order online at http://www.edpubs.org.

Content Contact
Eugene Owen
(202) 502-7422
Eugene.Owen@ed.gov

Executive Summary

The Program for International Student Assessment (PISA) is designed to monitor the mathematics, science, and reading achievement of students as they approach the end of compulsory schooling. PISA is sponsored by the Organization for Economic Cooperation and Development (OECD). The United States, other OECD member nations, and selected nonmember nations, participate in this program on a 3-year cycle. In each cycle, national samples of 15-year-old students in each participating country complete assessments in the specified subject areas and an associated questionnaire on related characteristics of the student. A companion questionnaire is completed by the principal of each sampled school. The intent is to establish how well students are acquiring the skills and knowledge necessary to participate fully in society now and in the future. The PISA 2003 data collection was fielded in April, May, and June 2003. To improve response rates (which were projected to be approximately 50 percent at the end of the original data collection period), a follow-up data collection was undertaken in August, September, and October 2003.

This report is concerned with the extent of potential bias introduced into the 2003 U.S. study through nonresponse on the part of schools and students. Specifically, the National Center for Education Statistics (NCES), which funded the U.S. study, requires that a nonresponse bias analysis be completed for any stage of data collection that has a unit response rate of less than 85 percent. In the 2003 U.S. PISA, the weighted school response rate (prior to the use of replacement schools) only reached 65 percent, substantially below the NCES-established standard. The weighted student response rate was 83 percent, also below the NCES-established standard.

Three bias analyses were conducted to address the issues of school nonresponse, student nonresponse, as well as possible achievement differences between students in the spring and fall testing sessions. Item-level nonresponse is not discussed in this report because response rates for all of the major reporting variables for this U.S. cycle (e.g., sex, socioeconomic status, and race/ethnicity) exceeded the NCES standard of 85 percent.

The analysis compares selected characteristics likely to reflect bias in participation from participating and nonparticipating schools. Frame characteristics for public schools were obtained from the 2000-01 Common Core of Data (CCD). For private schools, the characteristics were derived from the 2000-01 Private School Survey (PSS). The selected variables include school control, community type, National Assessment of Educational Progress (NAEP) region, poverty level, number of age-eligible students

enrolled, total number of students, and percentage of students by race/ethnicity categories.[1] The percentage of students eligible for the free or reduced-price lunch under the National School Lunch Program, available only for public schools, also was included.

Methodology

Two forms of analysis were undertaken:

- A test of the independence of each school and student characteristic and participation status, and
- A logistic regression in which the conditional independence of these school and student characteristics as predictors of participation was examined.

Significance tests used with analyses in the first of these categories were based on a Rao-Scott modified chi-square statistic in the case of categorical variables and a *t* test on the differences between means for continuous variables. The 95 percent confidence level was used for all tests of significance. All analyses were performed using WesVar (Westat 2002), with replicate weights to properly account for the complex sample design. Two sets of school weights were used: the school base weights that did not include a nonresponse adjustment factor, and nonresponse adjusted school weights.

The statistical analyses at the school level used two types of samples: the original and the final sample. The original sample treats all the schools that were replaced as nonparticipants; thus replacement schools are not included in this analysis. The final sample analysis includes the replacement schools in the participating schools group, and treats the nonparticipating schools that were not replaced as nonparticipants. The final sample analysis was repeated using the nonresponse adjusted school weights.

Effects of School Nonresponse

These analyses showed no statistically significant relationship between response status and the majority of school characteristics examined independently. However, three characteristics did show a significant relationship to response: NAEP region; percentage of Asian or Pacific Islander students; and percentage of Black, non-Hispanic students. Using the logistic regression model, none of the characteristics reached statistical significance. For the final sample of schools with school nonresponse adjustments applied to

[1] The definitions of the variables are found in section 2, Methodology.

the weights, there is no statistically significant relationship between participation and any available school characteristic.

Effects of Student Nonresponse

The basic form of analysis adopted to investigate student nonresponse was similar to that for school nonresponse, except that the focus was a comparison of participating students with all eligible sampled students in participating schools—6,803 students in all. The school-level variables extracted from the sampling frame were school control, community type, NAEP region, poverty level, and number of age-eligible students enrolled. The student variables were grade level, sex, and mean age.

A test of the independence of each school or student characteristic and student response status was performed along with a logistic regression designed to establish the conditional independence of these characteristics as predictors of student nonresponse. Two sets of student weights were used: the student base weights that did not include a nonresponse adjustment factor, and nonresponse adjusted student weights.

The findings suggest a statistically significant relationship between student response status and some of the school and student characteristics that were available for analysis. Specifically, school control, community type, NAEP region, number of age-eligible students, sex, and average age of students were related to response status. None of these differences appeared to be substantial, however.

Session Effects

Over the past decade or so, national data collection efforts that solicit the voluntary participation of schools have faced increasing resistance from school districts and schools. In 2003, about 47 percent of the original sample of schools agreed to participate in a spring assessment. However, close to 20 percent of refusing schools in the original sample cited time of year as the main problem and, when asked, agreed to undertake the assessment in the fall. With the permission of the PISA Consortium, the United States conducted additional assessments in the fall, adjusting the age eligibility definition so that all students tested in spring and fall 2003 were the same age. The intention was to combine the data from both spring and fall assessments if it could be shown that there was no significant between-session bias in the achievement results of the spring and fall samples.

Since schools, rather than students, opted to be assessed in either the spring or the fall, session effects were seen as school-level effects, with the potential to influence the assessment scores of all students within a school equally but varying between schools by session. This situation was modeled statistically as a two-level hierarchical linear model of the kind often identified as *means-as-outcomes* regression. In this model, the student-level model was fully unconditional with student scores a function of the school mean plus a random component. In the school-level model, session, variables of school control, community type, NAEP region, number of age-eligible students enrolled, and percentage of minority students were used to predict variation in school means. The analyses were based on data from 274 schools. Significance tests refer to the significance of partial regression coefficients at the 5 percent level of confidence.

Since the coefficient for the *session* variable failed to reach statistical significance, the analyses offered no support to the notion of session effect bias. On this basis, fall and spring 2003 PISA data for the United States were combined into a single dataset.

Acknowledgments

Many people contributed to making this report possible, and the authors wish to thank all those who have assisted with various aspects of the report.

Page intentionally left blank.

Contents

Contents

List of Tables

1. Introduction

The Program for International Student Assessment (PISA) is a large international comparative study of the knowledge, skills, and competencies of 15-year-old students in the domains of reading literacy, mathematics literacy, science literacy, and problem solving. The 2003 cycle of the study was carried out in some 41 countries, including the United States. To provide valid estimates of student achievement (and characteristics), the sample of PISA students was selected to represent the full population of 15-year-old students in each country. For PISA 2003, the international desired population in each country consisted of 15-year-olds attending educational institutions, both publicly and privately controlled, located within the country, in grades 7 and higher.

The U.S. PISA study, supported by the National Center for Education Statistics (NCES), utilized a two-stage stratified cluster sampling design. The first stage made use of a systematic probability-proportionate-to-size technique to select schools where size is the estimated age-eligible enrollment of the school. Though efforts were made to secure the participation of all schools selected in the first stage, it was anticipated that not all schools would choose to participate. Therefore, as each school was selected in the sample, the two neighboring schools in the sampling frame (immediately preceding and following it) were designated as replacement schools. If the sampled school refused to participate, the first replacement was then contacted. If that school also refused to participate, the second school was then contacted.

Within schools, a sample of 35 students was to be selected in an equal probability sample, unless fewer than 35 students aged 15 were available (in which case all students were selected). International standards required that students be sampled based on an age definition of 15 years and 3 months to 16 years and 2 months at the beginning of the testing period. The U.S. PISA sample consisted of 382 eligible schools[1] having at least one 15-year-old student. Detailed information on technical aspects of the study is included in appendix A.

The PISA data collection was fielded in April, May, and June 2003. In the United States, for a variety of reasons reported by school administrators (such as increased testing requirements at the national, state, and local levels; concerns about timing of the PISA assessment; and loss of learning time), many schools in the original sample declined to participate. As it was clear that the U.S. effort would not meet the minimum response rate standards required by NCES or the intermediate response rate zone required by the international standards (see appendix A for more details), a second testing window was opened from

[1] Of the 420 original schools selected for the sample, there were 38 ineligible or closed schools.

September to November 2003 with the agreement of the PISA Consortium in order to improve response rates and better accommodate school schedules. For the fall data collection, the school sample included only original schools from the sample that had refused to participate in the spring but indicated a willingness to participate in a fall assessment. Replacement schools were not included in the fall sample because their participation would have had little effect on raising the final response rate. In order to achieve a comparable sample of students in spring and fall, the age definition for students tested in the fall was adjusted such that all students tested were the same age. That is, in the spring sample 15-year-olds were defined as those born in the 1987 calendar year. In the case of the fall sample this age span was adjusted such that students born between June 1987 and May 1988 were defined as eligible.

Of the 420 schools sampled, 382 were determined to be eligible, and of these, 249 participated in either spring or fall 2003. For the purposes of calculating response rates, international guidelines stipulated that schools with less than 50 percent of students participating were considered not participating and their students were excluded from the student response rates. This resulted in 249 participating schools and an initial (before replacement) unweighted response rate of 65 percent (also 65 percent weighted). After recruitment of 13 replacement schools for a total of 262 participating schools (after replacement), the unweighted response rate increased to 69 percent (68 percent weighted).

Of the 9 initial schools with less than 50 percent of students participating, if the student response rates were at least 25 percent, these schools and students were included in the PISA 2003 database and the schools treated as participants in this report. Schools with student response rates below 25 percent were not used in any type of analysis nor are the data for these students or schools available in the PISA 2003 database, and the schools were treated as not participating in this report. The analysis of school nonresponse was thus based on 258 participating original and 274 final schools, respectively.[2] The school participation rates for this report are summarized in table 1-1. See appendix A and table A-1 for details on the U.S. PISA school participation rates.

[2] Three replacement schools with low response rates were also included in the final count of 274 schools used in this analysis.

Table 1-1. Selected characteristics for the nonresponse bias analysis of the U.S. PISA final school sample: 2003

Schools in original sample	Eligible schools in sample	Number of participating schools before replacement	Number of participating schools after replacement	Unweighted school participation rate before replacement (percent)	Weighted school participation rate before replacement (percent)	Unweighted school participation rate after replacement (percent)	Weighted school participation rate after replacement (percent)
420	382	258	274	67.5	67.4	71.7	71.4

SOURCE: Organization for Economic Cooperation and Development (OECD), Program for International Student Assessment (PISA), 2003.

With respect to student response, a total of 7,598 students were sampled for the assessment. Of these students, 261 were deemed ineligible because of their enrolled grades, birthdays, or other reasons and 534 students were excluded because of limited proficiency in English or functional or intellectual disabilities. They were all excluded from the response rates and this report. An additional 301 students (114 participating) came from schools with 25 to 49 percent student participation. They were excluded from the response rates as stipulated by international guidelines but included in this report, since their data are available in the PISA 2003 database. This resulted in a total of 5,456 participants of the 6,803 eligible students in this analysis for an unweighted student response rate of 80 percent (81 percent weighted). See appendix A and table A-2 for details on the U.S. PISA student participation rates.

NCES standards for assessment surveys stipulate that a nonresponse bias analysis is required at any stage of data collection reporting a weighted unit response rate less than 85 percent. Since the U.S. PISA weighted school and student response rates are below 85 percent, NCES requires an investigation into the potential magnitude of nonresponse bias at the school and student levels in the U.S. sample, which are the focus of this report. NCES also requested an investigation into the potential for bias that may have been introduced by the need to conduct the assessments in two sessions, one in the spring and the other in the fall.

This report is structured around these issues. Chapter 2 addresses the question of nonresponse bias among schools, chapter 3 focuses on the effects of student nonresponse, and chapter 4 investigates session effects. The methodology used to investigate nonresponse bias at each stage is included in each chapter. Conclusions follow in chapter 5. Item-level nonresponse is not discussed in this report, as none of the key variables in the U.S. survey (sex, socioeconomic status, and race/ethnicity) had response rates lower than the NCES standard of 85 percent.

2. School Nonresponse Bias

Methodology

To measure the potential nonresponse bias at the school level, the characteristics of participating schools were compared to those of the total eligible sample of schools. The alternative of comparing participants to nonparticipants through the use of the same tests of significance makes it more difficult to judge the potential for bias. The present analysis is similar to other National Center for Education Statistics (NCES) nonresponse bias studies on the 2005 National Assessment of Educational Progress (NAEP) (Kali and Rust 2005) and the 2001 Program for International Reading Literacy Study (PIRLS) (Piesse and Rust 2003).

The analysis for each grade was conducted in three parts:

- First, the distribution of the participating original school sample was compared with that of the total eligible original school sample. The original sample is the sample before substitution. In each sample, schools were weighted by their school base weights, excluding any nonresponse adjustment factor. The base weight for each original school was the reciprocal of its selection probability.

- Second, the distribution of the participating sample, which includes the participating replacements that were used as substitutes for schools from the original sample that did not participate, was compared to the total eligible final sample. The final sample is the sample after substitution. Again, school base weights were used for both the eligible sample and the participating schools. The base weight for each replacement school was set to the base weight of the original school that it replaced.

- Third, the same sets of schools were compared as in the second analysis but, this time, when analyzing the participating schools alone, school nonresponse adjustments were applied to the weights. The international weighting procedures formed nonresponse adjustment classes by cross-classifying the explicit and implicit stratification variables.

The first analysis indicates the potential for nonresponse bias that was introduced through school nonresponse. The second analysis suggests the remaining potential for nonresponse bias after the mitigating effects of substitution have been accounted for. The third analysis indicates the potential for bias after accounting for the mitigating effects of both substitution and nonresponse weight adjustments. Both the second and third analyses, however, may provide an overly optimistic scenario, resulting from the fact that substitution and nonresponse adjustments may correct somewhat for deficiencies in the characteristics examined here, but there is no guarantee that they are equally as effective for other characteristics and, in particular, for student achievement.

To compare Program for International Student Assessment (PISA) participants and the total eligible sample, the sample of schools was matched to the sample frame to compare as many characteristics as possible that might provide information about the presence of nonresponse bias. Comparing frame characteristics for participants and the total eligible sample is not an ideal measure of nonresponse bias if the characteristics are unrelated or weakly related to more substantive items in the survey; however, it often is the only approach available. While the school-level characteristics used in these analyses are limited to those available in the sampling frame, each of the variables has a demonstrated relationship to the achievement level of the school. Most, if not all, can be seen as indicators of school disadvantage with an established relationship to achievement (OECD 2004, pp. 255-268).

Frame characteristics for public schools were from the 2000-01 Common Core of Data (CCD) and, for private schools, from the 2000-01 Private School Survey (PSS).

The following categorical variables were available for all schools:

- School control—indicates whether the school is under public control (operated by publicly elected or appointed officials) or private control (operated by privately elected or appointed officials and derives its major source of funds from private sources);
- Community type—the location of a school relative to populous areas (i.e., central city, urban fringe/large town, rural/small town);
- NAEP region (see appendix A for state listing); and
- Poverty level[2]—for public schools, a high poverty school is defined as one in which 50 percent or more of the students are eligible for participation in the National School Lunch Program (NSLP), and a low poverty school is defined as one in which less than 50 percent are eligible; all private schools are treated as low poverty schools.

The following continuous variables were available for all schools:

- Number of grade 4 or grade 8 eligible students enrolled;
- Total number of students;
- Percentage White, non-Hispanic students enrolled in the school;
- Percentage Black, non-Hispanic students enrolled in the school;
- Percentage Hispanic students enrolled in the school;

[2] The sample frame did not contain a direct measure of poverty.

- Percentage Asian or Pacific Islander students enrolled in the school; and
- Percentage American Indian or Alaska Native students enrolled in the school.

An additional continuous variable, the percentage of students eligible to participate in the National School Lunch Program (NSLP), was available only for public schools.

For categorical variables, the distribution of frame characteristics for participants was compared with the distribution for all eligible schools. The hypothesis of independence between the characteristic and participation status was tested using a Rao-Scott modified Chi-square statistic at the 5 percent level (Rao and Thomas 2003). For continuous variables, summary means were calculated and the difference between means was tested using a *t* test. The *p*-values for the tests are presented in the tables. The statistical significance of differences between participants and the total eligible sample is identical to that which would result from comparing participants and nonparticipants, since all significance tests account for the fact that the participants are a subset of the full sample. The bias and relative bias are also shown in each table. The bias is the difference between the respective estimates for the participants and the eligible sample. The relative bias is calculated as the bias divided by the estimate from the eligible sample. The relative bias is a measure of the size of the bias compared to the eligible sample estimate.

In addition to these tests, logistic regression models were used to provide a multivariate analysis in which the conditional independence of these school characteristics as predictors of participation was examined. It may be that only one or two variables are actually related to participation status. However, if these variables are also related to the other variables examined in the analyses, then other variables, which are not related to participation status, will appear as significant in simple bivariate tables. Dummy variables were created for each component of the categorical variables so that each component was included separately. The last component of each categorical variable is used as the reference category and is not included in the model explicitly. The *p*-value of a dummy variable indicates whether there is a significant difference at the 5 percent level from the effect of the (omitted) reference category. All the frame characteristics were included in the model.

The logistic regression was performed using WesVar (Westat 2002) and replicate weights to properly account for the complex sample design. The Balanced Repeated Replication (BRR)—the Fay method of BRR was used to create the replicate weights (Brick, Morganstein, and Valliant 2000).

Original Respondent Sample

This section presents the results of the nonresponse bias analysis based exclusively on the original sample of 382 eligible U.S. schools for PISA. The distribution of the participating original school sample was compared with that of the total eligible original school sample using base weights in each case. The unweighted response rate was 68 percent, with 258 out of 382 eligible schools participating. The weighted response rate was 67 percent.

Categorical Variables

The distribution of participating and eligible schools in the PISA original sample by the four characteristics is shown in table 2-1. Based on these analyses, the chi-square statistic for NAEP region suggests that there is evidence of a relationship with participating in the assessment. There are no statistically significant relationships between participation status and any of the other characteristics shown in table 2-1.

Table 2-1. Percentage distribution of eligible and participating schools in the U.S. PISA original sample, by selected categorical variables: 2003

School characteristic	Sample schools: Eligible (percent) (*N*=382)	Sample schools: Participating (percent) (*N*=258)	Bias	Relative bias	Chi-square *p*-value
School control					0.198
Public	93.96	92.89	-1.07	-0.011	
Private	6.04	7.11	1.07	0.177	
Community type					0.442
Central city	30.97	30.39	-0.58	-0.019	
Urban fringe/large town	39.15	37.89	-1.26	-0.032	
Rural/small town	29.88	31.72	1.84	0.062	
NAEP region					0.003
Northeast	20.26	20.48	0.22	0.011	
Southeast	24.07	29.07	5.00	0.208	
Central	23.11	19.77	-3.34	-0.145	
West	32.56	30.68	-1.88	-0.058	
Poverty level					0.585
High	11.15	11.75	0.60	0.054	
Low	88.85	88.25	-0.60	-0.007	

NOTE: Detail may not sum to totals because of rounding. National Assessment of Educational Progress (NAEP) region is the state-based region of the country (see appendix for state listing.) For public schools, a high poverty school is defined as one in which 50 percent or more of the students are eligible for participation in the National School Lunch Program (NSLP); all private schools are treated as low poverty schools. Eligible schools have at least one 15-year-old student. Participating schools agreed to have their students assessed. The relative bias is calculated as the bias divided by the estimate from the eligible sample. Schools were weighted by their school base weights that did not include a nonresponse adjustment factor.

SOURCE: Organization for Economic Cooperation and Development (OECD), Program for International Student Assessment (PISA), 2003.

Continuous Variables

Summary means for each continuous variable for participating and eligible schools are shown in tables 2-2 through 2-4. Two schools had a missing value for the total school enrollment and 50 out of the 357 public schools had a missing value for the free or reduced-price lunch variable; these schools were excluded from the analysis.

There were no statistically significant differences between participating and eligible schools as shown in tables 2-2 through 2-4. However, this must be interpreted with caution for the free or reduced-price lunch variable because it is missing for 50 schools.

Table 2-2. Mean enrollment of eligible and participating schools in the U.S. PISA original sample: 2003

	Sample schools				
Student enrollment	Eligible (mean)	Participating (mean)	Bias	Relative bias	*t* test *p*-value
Total school	1,399.17[1]	1,372.13[2]	-27.04	-0.019	0.388
Age-eligible	362.36[3]	356.39[4]	-5.97	-0.016	0.509

[1] $N = 380$
[2] $N = 256$
[3] $N = 382$
[4] $N = 258$

NOTE: Information on total school enrollment is missing for two participating schools of the 382 eligible schools in the sample. Eligible schools have at least one 15-year-old student. Participating schools agreed to have their students assessed. The relative bias is calculated as the bias divided by the estimate from the eligible sample. Schools were weighted by their school base weights that did not include a nonresponse adjustment factor.
SOURCE: Organization for Economic Cooperation and Development (OECD), Program for International Student Assessment (PISA), 2003.

Table 2-3. Mean percentage of students in eligible and participating schools in the U.S. PISA original sample, by race/ethnicity: 2003

	Sample schools				
Race/ethnicity	Eligible (percent) ($N = 382$)	Participating (percent) ($N = 258$)	Bias	Relative bias	*t* test *p*-value
White, non-Hispanic	62.51	61.14	-1.37	-0.022	0.191
Black, non-Hispanic	15.38	16.83	1.45	0.094	0.111
Hispanic	14.56	14.42	-0.14	-0.010	0.858
Asian or Pacific Islander	4.40	3.82	-0.58	-0.132	0.099
American Indian or Alaska Native	0.94	0.99	0.05	0.053	0.723

NOTE: Black includes African American, and Hispanic includes Latino. Racial categories exclude Hispanic origin. Eligible schools have at least one 15-year-old student. Participating schools agreed to have their students assessed. The relative bias is calculated as the bias divided by the estimate from the eligible sample. Schools were weighted by their school base weights that did not include a nonresponse adjustment factor.
SOURCE: Organization for Economic Cooperation and Development (OECD), Program for International Student Assessment (PISA), 2003.

Table 2-4. Mean percentage of students eligible for free or reduced-price lunch, in eligible and participating public schools in the U.S. PISA original sample: 2003

Students	Sample schools: Eligible (percent) (*N*=307)	Sample schools: Participating (percent) (*N*=205)	Bias	Relative bias	*t* test *p*-value
Percentage of students eligible for free or reduced-price lunch	26.56	28.01	1.45	0.055[!]	0.084[!]

! Interpret data with caution.
NOTE: Information on percentage of students eligible for free or reduced-price lunch is missing for 50 of the 357 public schools in the sample. Eligible schools have at least one 15-year-old student. Participating schools agreed to have their students assessed. The relative bias is calculated as the bias divided by the estimate from the eligible sample. Schools were weighted by their school base weights that did not include a nonresponse adjustment factor.
SOURCE: Organization for Economic Cooperation and Development (OECD), Program for International Student Assessment (PISA), 2003.

Logistic Regression Model

To examine the joint relationship of various characteristics to school nonresponse, the analysis utilized a logistic regression model with participation status as the binary dependent variable and frame characteristics as predictor variables. Public and private schools were modeled together using the variables available for all schools. Two schools were excluded from the analysis due to missing information for total school enrollment.

Standard errors and tests of hypotheses for the full model parameter estimates are presented in table 2-5. None of the parameter estimates are significant, which indicates that there are no significant relationships with participation status.

Table 2-5. Logistic regression model parameters using the U.S. PISA original school sample: 2003

Parameter	Parameter estimate	Standard error	*t* test for H_0: parameter = 0	*p*-value
Intercept	1.23	1.159	1.059	0.293
Private school	0.92	0.685	1.343	0.183
Central city	-0.17	0.268	-0.638	0.525
Rural/small town	0.24	0.300	0.812	0.419
Northeast	0.13	0.364	0.351	0.727
Southeast	0.61	0.462	1.330	0.187
Central	-0.36	0.348	-1.032	0.305
High poverty	0.03	0.451	0.070	0.945
Total school enrollment	#	0.001	-0.441	0.660
Age-eligible enrollment	#	0.002	0.485	0.629
White, non-Hispanic	-0.01	0.012	-0.615	0.540
Black, non-Hispanic	0.00	0.013	0.063	0.950
Hispanic	0.00	0.013	-0.331	0.741
Asian or Pacific Islander	-0.02	0.019	-1.147	0.255
American Indian or Alaska Native	0.01	0.022	0.276	0.783

Rounds to zero.

NOTE: Two of the 382 eligible schools in the sample were excluded due to missing information for total school enrollment. NAEP region is the state-based region of the country (see appendix for state listing.) For public schools, a high poverty school is defined as one in which 50 percent or more of the students are eligible for participation in the National School Lunch Program (NSLP); all private schools are treated as low poverty schools. Black includes African American, and Hispanic includes Latino. Racial categories exclude Hispanic origin. Schools were weighted by their school base weights that did not include a nonresponse adjustment factor.

SOURCE: Organization for Economic Cooperation and Development (OECD), Program for International Student Assessment (PISA), 2003.

Respondent Sample with Replacements (Final Sample)

This section presents the nonresponse bias analysis based on the final sample of 382 eligible schools for PISA. The distribution of the participating sample, including participating replacements, was compared to the total eligible final sample. School base weights were used for both the eligible sample and the participating schools. Through the use of replacements, the unweighted school response rate for PISA was 72 percent, with 274 out of 382 schools participating. The weighted response rate was 71 percent.

Categorical Variables

The distribution of participating and eligible schools by the four characteristics is shown in table 2-6.

Based on these analyses, the chi-square statistic for NAEP region suggests that there is evidence of a relationship with participating in the assessment. NAEP region was also significant in the original sample (table 2-1). While the use of replacement schools substantially reduced the bias in the Southeast (5.00 vs.

3.75) and Central (-3.34 vs. -1.50) regions, the bias increased in the Northeast (0.22 vs. 0.96) and West (-1.88 vs. -3.21). Thus, while there is certainly no evidence that the use of replacement schools reduced the potential for bias as indicated by this variable, it has also not substantially added to it. There are no statistically significant relationships between participation status and any of the other characteristics shown in table 2-6.

Table 2-6. Percentage distribution of eligible and participating schools in the U.S. PISA final sample, by selected categorical variables: 2003

School characteristic	Sample schools: Eligible (percent) (*N*=382)	Sample schools: Participating (percent) (*N*=274)	Bias	Relative bias	Chi-square *p*-value
School control					0.226
Public	93.96	93.04	-0.92	-0.010	
Private	6.04	6.96	0.92	0.152	
Community type					0.097
Central city	30.97	29.83	-1.14	-0.037	
Urban fringe/large town	39.15	37.66	-1.49	-0.038	
Rural/small town	29.88	32.51	2.63	0.088	
NAEP region					0.011
Northeast	20.26	21.22	0.96	0.047	
Southeast	24.07	27.82	3.75	0.156	
Central	23.11	21.61	-1.50	-0.065	
West	32.56	29.35	-3.21	-0.099	
Poverty level					0.872
High	11.69	11.85	0.16	0.014	
Low	88.31	88.15	-0.16	-0.002	

NOTE: Detail may not sum to totals because of rounding. National Assessment of Educational Progress (NAEP) region is the state-based region of the country (see appendix for state listing.) For public schools, a high poverty school is defined as one in which 50 percent or more of the students are eligible for participation in the National School Lunch Program (NSLP); all private schools are treated as low poverty schools. Eligible schools have at least one 15-year-old student. Participating schools agreed to have their students assessed. The relative bias is calculated as the bias divided by the estimate from the eligible sample. Schools were weighted by their school base weights that did not include a nonresponse adjustment factor.

SOURCE: Organization for Economic Cooperation and Development (OECD), Program for International Student Assessment (PISA), 2003.

Continuous Variables

Summary means for each continuous variable for participating and eligible schools are shown in tables 2-7 through 2-9. Two schools had a missing value for the total school enrollment, and 51 out of the 357 public schools had a missing value for the free or reduced-price lunch variable; these schools were excluded from the analysis.

Table 2-7. Mean enrollment of eligible and participating schools in the U.S. PISA final sample: 2003

Student enrollment	Sample schools: Eligible (mean)	Sample schools: Participating (mean)	Bias	Relative bias	*t* test *p*-value
Total school	1,398.55[1]	1,362.41[2]	-36.14	-0.026	0.205
Age-eligible	362.59[3]	354.19[4]	-8.40	-0.023	0.304

[1] *N*=380
[2] *N*=272
[3] *N*=382
[4] *N*=274

NOTE: Information on total school enrollment is missing for two participating schools of the 382 eligible schools in the sample. Eligible schools have at least one 15-year-old student. Participating schools agreed to have their students assessed. The relative bias is calculated as the bias divided by the estimate from the eligible sample. Schools were weighted by their school base weights that did not include a nonresponse adjustment factor.
SOURCE: Organization for Economic Cooperation and Development (OECD), Program for International Student Assessment (PISA), 2003.

Table 2-8. Mean percentage of students in eligible and participating schools in the U.S. PISA final sample, by race/ethnicity: 2003

Race/ethnicity	Sample schools: Eligible (percent) (*N*=382)	Sample schools: Participating (percent) (*N*=274)	Bias	Relative bias	*t* test *p*-value
White, non-Hispanic	62.39	61.64	-0.75	-0.012	0.412
Black, non-Hispanic	15.51	17.10	1.59	0.103	0.030
Hispanic	14.55	13.91	-0.64	-0.044	0.397
Asian or Pacific Islander	4.40	3.74	-0.66	-0.150	0.050
American Indian or Alaska Native	0.94	0.98	0.04	0.043	0.812

NOTE: Black includes African American, and Hispanic includes Latino. Racial categories exclude Hispanic origin. Eligible schools have at least one 15-year-old student. Participating schools agreed to have their students assessed. The relative bias is calculated as the bias divided by the estimate from the eligible sample. Schools were weighted by their school base weights that did not include a nonresponse adjustment factor.
SOURCE: Organization for Economic Cooperation and Development (OECD), Program for International Student Assessment (PISA), 2003.

Table 2-9. Mean percentage of students eligible for free or reduced-price lunch, in eligible and participating public schools in the U.S. PISA final sample: 2003

Students	Sample schools: Eligible (percent) (*N*=306)	Sample schools: Participating (percent) (*N*=217)	Bias	Relative bias	*t* test *p*-value
Percentage of students eligible for free or reduced-price lunch	26.84	28.06	1.22	0.045!	0.104!

! Interpret data with caution.
NOTE: Information on percentage of students eligible for free or reduced-price lunch is missing for 51 of the 357 public schools in the sample. Eligible schools have at least one 15-year-old student. Participating schools agreed to have their students assessed. The relative bias is calculated as the bias divided by the estimate from the eligible sample. Schools were weighted by their school base weights that did not include a nonresponse adjustment factor.
SOURCE: Organization for Economic Cooperation and Development (OECD), Program for International Student Assessment (PISA), 2003.

There were no statistically significant enrollment differences between participating and eligible schools (table 2-7). Participating schools had a higher mean percentage of Black, non-Hispanic students (17.1 vs. 15.5 percent, respectively; table 2-8) and lower of Asian or Pacific Islander students (3.7 vs. 4.4 percent, respectively; table 2-8) than the eligible sample. The difference in the percentage of students who are Black or Asian or Pacific Islander was small in absolute bias but substantial in relative bias. Therefore, there is some potential for bias existing from the use of replacement schools since it did not exist when looking only at the original sample schools (table 2-3). Thus, while there is certainly no evidence that the use of replacement schools reduced the potential for bias as indicated by this variable, it has also not substantially added to it as the change in relative bias is less than 1 percent (0.094 vs. 0.103) and 2 percent (0.132 vs. 0.150) for Blacks and Asian or Pacific Islander, respectively. The differences in the mean percentage of students of the other races and ethnicities (White, Hispanic, and American Indian or Alaska Native) were not statistically significant between the participating and eligible schools (table 2-8). There was no statistically significant difference between participating and eligible schools for percentage of students eligible for free or reduced-price lunch (table 2-9). However, this must be interpreted with caution for the free or reduced-price lunch variable because it is missing for 51 schools.

Logistic Regression Model

To examine the joint relationship of various characteristics to school nonresponse, the analysis utilized a logistic regression model with participation status as the binary dependent variable and frame characteristics as predictor variables. Two schools were excluded from the analysis due to missing information for total school enrollment.

Standard errors and tests of hypotheses for the full model parameter estimates are presented in table 2-10. None of the parameter estimates are significant, which indicates that there are no significant relationships with participation status.

Table 2-10. Logistic regression model parameters using the U.S. PISA final school sample: 2003

Parameter	Parameter estimate	Standard error	*t* test for H_0: parameter = 0	*p*-value
Intercept	1.28	1.132	1.130	0.262
Private school	1.02	0.767	1.333	0.186
Central city	-0.24	0.282	-0.861	0.392
Rural/small town	0.46	0.272	1.708	0.091
Northeast	0.36	0.363	1.001	0.320
Southeast	0.46	0.475	0.977	0.331
Central	-0.12	0.376	-0.316	0.752
High poverty	-0.15	0.496	-0.296	0.768
Total school enrollment	#	0.001	-0.426	0.671
Age-eligible enrollment	#	0.002	0.524	0.602
White, non-Hispanic	-0.01	0.012	-0.688	0.494
Black, non-Hispanic	0.01	0.013	0.479	0.633
Hispanic	-0.01	0.013	-0.384	0.702
Asian or Pacific Islander	-0.03	0.019	-1.311	0.194
American Indian or Alaska Native	0.01	0.023	0.240	0.811

Rounds to zero.
NOTE: Two of the 382 eligible schools in the sample were excluded due to missing information for total school enrollment. National Assessment of Educational Progress (NAEP) region is the state-based region of the country (see appendix for state listing.) For public schools, a high poverty school is defined as one in which 50 percent or more of the students are eligible for participation in the National School Lunch Program (NSLP); all private schools are treated as low poverty schools. Black includes African American, and Hispanic includes Latino. Racial categories exclude Hispanic origin. Schools were weighted by their school base weights that did not include a nonresponse adjustment factor.
SOURCE: Organization for Economic Cooperation and Development (OECD), Program for International Student Assessment (PISA), 2003.

Final Sample, With Nonresponse Adjustments Applied

This section presents the nonresponse bias analysis based on the final sample of 382 schools for PISA. These are the same sets of schools that were compared as in the previous analysis, but this time when analyzing the participating schools, school nonresponse adjustments were applied to the weights.

Categorical Variables

The distribution of participating and eligible schools by the four characteristics is shown in table 2-11. There are no statistically significant relationships between participation status and any of the characteristics, shown in table 2-11.

Table 2-11. Percentage distribution of eligible and participating schools in the U.S. PISA final sample after nonresponse adjustment, by selected categorical variables: 2003

School characteristic	Sample schools: Eligible (percent) (*N*=382)	Sample schools: Participating (percent) (*N*=274)	Bias	Relative bias	Chi-square *p*-value
School control					0.162
Public	93.96	93.95	-0.01	0.000	
Private	6.04	6.05	0.01	0.002	
Community type					0.453
Central city	30.97	31.86	0.89	0.029	
Urban fringe/large town	39.15	38.27	-0.88	-0.022	
Rural/small town	29.88	29.87	-0.01	0.000	
NAEP region					0.178
Northeast	20.26	21.30	1.04	0.051	
Southeast	24.07	23.64	-0.43	-0.018	
Central	23.11	22.70	-0.41	-0.018	
West	32.56	32.36	-0.20	-0.006	
Poverty level					0.675
High	11.69	12.11	0.42	0.036	
Low	88.31	87.89	-0.42	-0.005	

NOTE: Detail may not sum to totals because of rounding. For public schools, a high poverty school is defined as one in which 50 percent or more of the students are eligible for participation in the National School Lunch Program (NSLP); all private schools are treated as low poverty schools. National Assessment of Educational Progress (NAEP) region is the state-based region of the country (see appendix for state listing.) Eligible schools have at least one 15-year-old student. Participating schools agreed to have their students assessed. The relative bias is calculated as the bias divided by the estimate from the eligible sample. Schools were weighted by their school nonresponse adjusted weights.
SOURCE: Organization for Economic Cooperation and Development (OECD), Program for International Student Assessment (PISA), 2003.

Continuous Variables

Summary means for each continuous variable for participating and eligible schools are shown in tables 2-12 through 2-14. Two schools had a missing value for the total school enrollment, and 51 out of the 357 public schools had a missing value for the free or reduced-price lunch variable; these schools were excluded from the analysis.

Based on the analyses shown in tables 2-12 through 2-14, there were no statistically significant differences between participating and eligible schools. However, this must be interpreted with caution for the free or and reduced-price lunch variable because it is missing for 51 schools.

Table 2-12. Mean enrollment of eligible and participating schools in the U.S. PISA final sample after nonresponse adjustment: 2003

	Sample schools				
Student enrollment	Eligible (mean)	Participating (mean)	Bias	Relative bias	*t* test *p*-value
Total school	1,398.55[1]	1,395.51[2]	-3.04	-0.002	0.904
Age-eligible	362.59[3]	364.85[4]	2.26	0.006	0.749

[1] *N*=380
[2] *N*=272
[3] *N*=382
[4] *N*=274

NOTE: Information on total school enrollment is missing for two participating schools of the 382 eligible schools in the sample. Eligible schools have at least one 15-year-old student. Participating schools agreed to have their students assessed. The relative bias is calculated as the bias divided by the estimate from the eligible sample. Schools were weighted by their school nonresponse adjusted weights.
SOURCE: Organization for Economic Cooperation and Development (OECD), Program for International Student Assessment (PISA), 2003.

Table 2-13. Mean percentage of students in eligible and participating schools in the U.S. PISA final sample after nonresponse adjustment, by race/ethnicity: 2003

	Sample schools				
Race/ethnicity	Eligible (percent) (*N*=382)	Participating (percent) (*N*=274)	Bias	Relative bias	*t* test *p*-value
White, non-Hispanic	62.39	61.31	-1.08	-0.017	0.263
Black, non-Hispanic	15.51	16.75	1.24	0.080	0.101
Hispanic	14.55	14.69	0.14	0.010	0.833
Asian or Pacific Islander	4.40	3.97	-0.43	-0.098	0.228
American Indian or Alaska Native	0.94	1.06	0.12	0.128	0.545

NOTE: Black includes African American, and Hispanic includes Latino. Racial categories exclude Hispanic origin. Eligible schools have at least one 15-year-old student. Participating schools agreed to have their students assessed. The relative bias is calculated as the bias divided by the estimate from the eligible sample. Schools were weighted by their school nonresponse adjusted weights.
SOURCE: Organization for Economic Cooperation and Development (OECD), Program for International Student Assessment (PISA), 2003.

Table 2-14. Mean percentage of students eligible for free or reduced-price lunch, in eligible and participating public schools in the U.S. PISA final sample after nonresponse adjustment: 2003

Students	Sample schools: Eligible (percent) (N = 306)	Sample schools: Participating (percent) (N = 217)	Bias	Relative bias	t test p-value
Percentage of students eligible for free or reduced-price lunch	26.84	27.80	0.96	0.036	0.183

NOTE: Information on percentage of students eligible for free or reduced-price lunch is missing for 51 of the 357 public schools in the sample. Eligible schools have at least one 15-year-old student. Participating schools agreed to have their students assessed. The relative bias is calculated as the bias divided by the estimate from the eligible sample. Schools were weighted by their school nonresponse adjusted weights.
SOURCE: Organization for Economic Cooperation and Development (OECD), Program for International Student Assessment (PISA), 2003.

Logistic Regression Model

The analysis of the final sample of original and replacement schools utilized two logistic regression models both with participation status as the binary dependent variable. The first model had frame characteristics (as in table 2-10) plus school nonresponse adjustment class as predictor variables to examine the relationship of the school nonresponse class to school participation. The second model only had school nonresponse adjustment class as a predictor variable. In general, nonresponse adjustment classes are formed based on characteristics related to response rates or to values of survey estimates where respondents and nonrespondents are similar within each class. The nonresponse adjustment is applied within each of these classes. The international weighting procedures formed nonresponse adjustment classes by cross-classifying the explicit and implicit stratification variables. This resulted in 39 classes after collapsing. For the first model, an F test was used to determine whether the frame characteristics, in aggregate, were significantly related to response, once the nonresponse adjustment classes were taken into account. The F test statistic is 0.99 with a p-value of 0.448, which indicates no significant relationship with participation. The second model used an F test to determine whether the nonresponse adjustment class variable was significantly related to school response. The F test statistic is 7.45 with a p-value of less than 0.001, which indicates a significant relationship with school participation.

Summary

The investigation into nonresponse bias at the school level for the U.S. PISA effort shows no statistically significant relationship between response status and the majority of the available school characteristics that were examined in the analysis.

For original sample schools, only NAEP region was found to be significantly different between the participating schools and the total eligible sample of schools when looking at the test for independence. However, the regression analysis does not show any variables to be significant predictors of participation.

For final sample schools, NAEP region was again found to be significantly different between the participating schools and the total eligible sample of schools when looking at the test for independence. Additionally, there was a difference found in the percentage of Black, non-Hispanic and Asian or Pacific Islander students. Again however, the regression analysis does not show any variables to be significant predictors of school participation.

For the final sample of schools with school nonresponse adjustments applied to the weights, there is no statistically significant relationship between participation and any available school characteristic in the bivariate analyses. School nonresponse adjustment class was a significant predictor of school participation in the logistic regression. The school nonresponse adjustments were effective in reducing the bias in the characteristics studied.

3. Student Nonresponse Bias

This section presents the results of the student nonresponse bias analysis based exclusively on the sample of 6,803 eligible U.S. students for the Program for International Student Assessment (PISA). The distribution of the participating student sample was compared with that of the total eligible student sample using student base weights in each case. The unweighted response rate was 80 percent, with 5,456 out of 6,803 eligible students participating. The weighted response rate was 81 percent. See table A-2 for details on the PISA student participation rates.

Methodology

To measure the potential nonresponse bias at the student level, as was the case at the school level, the characteristics of participating students were compared to those of the total eligible sample of students. The alternative of comparing participants to nonparticipants, while resulting in the same tests of significance, makes it more difficult to judge the potential for bias. This analysis is similar to other National Center for Education Statistics (NCES) nonresponse bias studies on the 2005 National Assessment of Educational Progress (NAEP) (Kali and Rust 2005).

The analysis was conducted in two parts:

- First, the distribution of the participating student sample was compared with that of the total eligible student sample. Students were weighted by their student base weights that did not include a nonresponse adjustment factor. The base weight for each student was the reciprocal of the student within school selection probability times the final school weight.
- Second, the same sets of students were compared as in the first analysis, but this time, when analyzing the participating students alone, student nonresponse adjustments were applied to the weights. The international weighting procedures created nonresponse adjustment classes within school by high/low grade combination.[3]

The first analysis indicates the potential for nonresponse bias that was introduced through student nonresponse. The second analysis indicates the potential for bias after accounting for the mitigating effects of nonresponse weight adjustments. The second analyses, however, may provide an overly optimistic scenario since nonresponse adjustments may correct somewhat for deficiencies in the few

[3] The high/low grade classes in each country were defined so as to each contain a substantial proportion of the PISA population. For the United States, the classes were grades 9 and below and grades 10 and above.

characteristics examined here but without any guarantee that they are equally as effective for other characteristics and, in particular, for student achievement.

This analysis was conducted on two sets of characteristics: school level, from the sampling frame, and student level, collected on all students during the assessment. To compare PISA participants and the total eligible sample on the school-level characteristics, it was necessary to match the sample of schools back to the sample frame to select as many school characteristics as possible that might provide information about the presence of nonresponse bias. Comparing frame characteristics for participants and the total eligible sample is not an ideal measure of nonresponse bias if the characteristics are unrelated or weakly related to more substantive items in the survey; however, this is often the only approach available. With three exceptions, the student-level characteristics available for these analyses can be considered as contextual influences on student performance. The three exceptions are grade, sex, and age, which are individual student characteristics. Achievement differences by grade and sex are well established in the literature, and since age and grade are confounded in this sample, one would also expect a relationship between age and achievement.

Frame characteristics were from the 2000-01 Common Core of Data (CCD) for public schools, and from the 2000-01 Private School Survey (PSS) for private schools. The following school-level variables were available for all schools:

- School control—indicates whether the school is under public control (operated by publicly elected or appointed officials) or private control (operated by privately elected or appointed officials and derives its major source of funds from private sources);
- Community type—the location of a school relative to populous areas (i.e., central city, urban fringe/large town, rural/small town);
- NAEP region (see appendix for state listing);
- Poverty level[3]—for public schools, a high poverty school is defined as one in which 50 percent or more of the students are eligible for participation in the National School Lunch Program (NSLP); all private schools are treated as low poverty schools; and
- Number of age-eligible students enrolled.

[3] The sample frame did not contain a direct measure of poverty.

The following student-level characteristics were composed of the limited number that was collected on all eligible students at the assessment:

- Grade level—7th through 11th grade with the 7th and 8th grade students collapsed into a single cell due to the small number of 7th grade students;
- Sex; and
- Average age of student (in months).

For categorical variables, the distribution of frame characteristics for participants was compared with the distribution for all eligible students. The hypothesis of independence between the characteristic and participation status was tested using a Rao-Scott modified chi-square statistic at the 5 percent level (Rao and Thomas 2003). For continuous variables, summary means were calculated and the difference between means was tested using a *t* test. The *p*-values for the tests are presented in the tables. School control, community type, NAEP region, poverty grade level and sex are categorical variables, and the number of age-eligible students enrolled and average age of student are continuous. The statistical significance of differences between participants and the total eligible sample is identical to that which would result from comparing participants and nonparticipants since all significance tests account for the fact that the participants are a subset of the full sample. The bias and relative bias are also given in each table. The bias is the difference between the respective estimates for the participants and the eligible sample. The relative bias is calculated as the bias divided by the estimate from the eligible sample. The relative bias is a measure of the size of the bias compared to the eligible sample estimate.

In addition to these tests, logistic regression models were used to provide a multivariate analysis in which the conditional independence of these school or student characteristics as predictors of participation was examined. It may be that only one or two variables are actually related to participation status. However, if these variables are also related to the other variables examined in the analyses, then other variables, which are not related to participation status, will appear as significant in simple bivariate tables. Dummy variables were created for each component of the categorical variables so that each component was included separately. The last component of each categorical variable is always the reference category and is not included in the model explicitly. The *p*-value of a dummy variable indicates whether there is a significant difference at the 5 percent level from the effect of the (omitted) reference category. All the characteristics were included in the model.

The logistic regression was performed using WesVar (Westat 2002) and replicate weights to properly account for the complex sample design. The Balanced Repeated Replication (BRR)—the Fay method of BRR—was used to create the replicate weights (Brick, Morganstein, and Valliant 2000).

Student Respondent Sample, With Base Weights

The distribution of the participating sample was compared to the total eligible sample. Student base weights were used for both the eligible sample and the participating students. The unweighted student response rate for PISA was 80 percent, with 5,456 out of 6,803 eligible students participating. The weighted response rate was 81 percent.

School-level Variables

The distribution of participating and eligible students by the four characteristics is shown in table 3-1. Based on these analyses, the chi-square statistic for school control suggests that there is evidence of a relationship with participating in the assessment. Specifically, students in private schools participate at a higher level than students in public schools. The chi-square statistic for community type and NAEP region also suggests that there is evidence of a relationship with participating in the assessment. There are no statistically significant relationships between participation status and poverty level.

The mean number of age-eligible students for participating and eligible students is shown in table 3-2. The *t* test for the mean number of age-eligible students indicates that participating students were in schools with fewer age-eligible students than the eligible sample.

Table 3-1. Percentage distribution of eligible and participating students in the U.S. PISA student sample, by selected school-level categorical variables: 2003

School characteristic	Sample students: Eligible (percent) (*N*=6,803)	Sample students: Participating (percent) (*N*=5,456)	Bias	Relative bias	Chi-square *p*-value
School control					<0.000
Public	92.77	91.80	-0.97	-0.010	
Private	7.23	8.20	0.97	0.134	
Community type					0.001
Central city	28.08	26.78	-1.30	-0.046	
Urban fringe/large town	38.74	38.10	-0.64	-0.017	
Rural/small town	33.18	35.12	1.94	0.058	
NAEP region					0.006
Northeast	21.02	20.23	-0.79	-0.038	
Southeast	29.26	31.20	1.94	0.066	
Central	21.78	21.00	-0.78	-0.036	
West	27.94	27.56	-0.38	-0.014	
Poverty level					0.646
High	11.25	11.11	-0.14	-0.012	
Low	88.75	88.89	0.14	0.002	

NOTE: Detail may not sum to totals because of rounding. National Assessment of Educational Progress (NAEP) region is the state-based region of the country (see appendix for state listing.) For public schools, a high poverty school is defined as one in which 50 percent or more of the students are eligible for participation in the National School Lunch Program (NSLP); all private schools are treated as low poverty schools. Eligible students were 15 years old as defined by the Program for International Student Assessment (PISA). Participating students were those assessed. The relative bias is calculated as the bias divided by the estimate from the eligible sample. Students were weighted by their student base weights that did not include a nonresponse adjustment factor.
SOURCE: Organization for Economic Cooperation and Development (OECD), Program for International Student Assessment (PISA), 2003.

Table 3-2. Mean number of age-eligible students in the U.S. PISA student sample: 2003

Student enrollment	Sample students: Eligible (mean) (*N*=6,803)	Sample students: Participating (mean) (*N*=5,456)	Bias	Relative bias	*t* test *p*-value
Age-eligible	329.47	318.01	-11.46	-0.035	<0.000

NOTE: Eligible students were 15 years old as defined by the Program for International Student Assessment (PISA). Participating students were those assessed. The relative bias is calculated as the bias divided by the estimate from the eligible sample. Students were weighted by their student base weights that did not include a nonresponse adjustment factor.
SOURCE: Organization for Economic Cooperation and Development (OECD), Program for International Student Assessment (PISA), 2003.

Student-level Variables

The distribution of participating and eligible students by the two characteristics is shown in table 3-3. Eleven students were missing values for grade level, and two students were missing values for sex and were not included in the analysis. The chi-square statistic for sex indicates females participate at a higher level than males. There are no statistically significant relationships between participation status and grade.

The mean age of students (in months) for participating and eligible students is shown in table 3-4. Age was missing for one student and was excluded from the analysis. The *t* test for the mean age of students indicates that the average age of participating students was younger than the eligible sample. However, the size of the difference was no more than three days.

Table 3-3. Percentage distribution of eligible and participating students in the U.S. PISA student sample, by selected student-level categorical variables: 2003

Student characteristic	Sample students: Eligible (percent)	Sample students: Participating (percent)	Bias	Relative bias	Chi-square *p*-value
Grade[1,2]					0.611
7th and 8th	3.02	3.10	0.08	0.026	
9th	29.83	29.98	0.15	0.005	
10th	60.44	60.44	#	#	
11th	6.72	6.48	-0.24	-0.036	
Sex[3,4]					0.037
Female	49.36	50.02	0.66	0.013	
Male	50.64	49.98	-0.66	-0.013	

Rounds to zero.
[1] $N = 6{,}792$ for eligible
[2] $N = 5{,}454$ for participating
[3] $N = 6{,}801$ for eligible
[4] $N = 5{,}455$ for participating
NOTE: Information on grade level is missing for 11 of the 6,803 eligible students. Information on sex is missing for two of the 6,803 eligible students. Eligible students were 15-years-old as defined by the Program for International Student Assessment (PISA). Participating students were those assessed. The relative bias is calculated as the bias divided by the estimate from the eligible sample. Students were weighted by their student base weights that did not include a nonresponse adjustment factor.
SOURCE: Organization for Economic Cooperation and Development (OECD), Program for International Student Assessment (PISA), 2003.

Table 3-4. Mean age of eligible and participating students in the U.S. PISA student sample: 2003

Student age (months)	Sample students: Eligible (mean) (*N*=6,802)	Sample students: Participating (mean) (*N*=5,456)	Bias	Relative bias	*t* test *p*-value
Age	190.05	189.98	-0.07	#	0.007

Rounds to zero.
NOTE: Information on age is missing for one of the 6,803 eligible students. Eligible students were 15 years old as defined by the Program for International Student Assessment (PISA). Participating students were those assessed. The relative bias is calculated as the bias divided by the estimate from the eligible sample. Students were weighted by their student base weights that did not include a nonresponse adjustment factor.
SOURCE: Organization for Economic Cooperation and Development (OECD), Program for International Student Assessment (PISA), 2003.

Logistic Regression Model

To examine the joint relationship of various characteristics to student nonresponse, the analysis utilized a logistic regression model with participation status as the binary dependent variable and frame characteristics as predictor variables. Fourteen students were excluded due to missing information for grade, sex, or age.

Standard errors and tests of hypotheses for the full model parameter estimates are presented in table 3-5. Private schools, Northeast region, Central region, age-eligible enrollment, female and age are significant predictors of student participation of the characteristics examined. The positive parameter estimate for private schools indicates that students in private schools were more likely to participate than students in public schools. The negative parameter estimates for the Northeast and Central regions indicate that students in those regions were less likely to participate than students in the West region. The negative parameter estimate for age-eligible enrollment indicates that students in smaller schools were more likely to participate than students in larger schools. The negative parameter estimate for age indicates that students who participated tend to be younger than the eligible sample.

Table 3-5. Logistic regression model parameters using the U.S. PISA student sample: 2003

Parameter	Parameter estimate	Standard error	*t* test for H_0: parameter = 0	*p*-value
Intercept	8.10	2.405	3.367	0.001
Private school	1.04	0.387	2.672	0.009
Central city	-0.25	0.164	-1.537	0.128
Rural/small town	0.24	0.160	1.507	0.136
Northeast	-0.40	0.196	-2.044	0.044
Southeast	0.28	0.163	1.687	0.096
Central	-0.34	0.169	-2.006	0.048
High poverty	0.07	0.186	0.372	0.711
Age-eligible	#	#	-2.997	0.004
Grade 7th and 8th	-0.31	0.284	-1.073	0.286
Grade 9th	-0.06	0.231	-0.250	0.803
Grade 10th	0.01	0.192	0.052	0.959
Female	0.13	0.063	2.040	0.045
Age (month)	-0.03	0.012	-2.736	0.008

Rounds to zero.

NOTE: Fourteen of the 6,803 eligible students in the sample were excluded due to missing information for grade, sex or age. National Assessment of Educational Progress (NAEP) region is the state-based region of the country (see appendix for state listing.) For public schools, a high poverty school is defined as one in which 50 percent or more of the students are eligible for participation in the National School Lunch Program (NSLP); all private schools are treated as low poverty schools. Students were weighted by their school base weights that did not include a nonresponse adjustment factor.

SOURCE: Organization for Economic Cooperation and Development (OECD), Program for International Student Assessment (PISA), 2003.

Respondent Sample, With Nonresponse Adjustments Applied

This section presents the nonresponse bias analysis based on the final eligible sample of 6,803 students for PISA. This is the same sets of students that were compared as in the previous analysis, but this time when analyzing the participating students alone, student nonresponse adjustments were applied to the weights.

School-level Variables

The distribution of participating and eligible students by the four characteristics is shown in table 3-6.

Table 3-6. Percentage distribution of eligible and participating students in the U.S. PISA student sample after nonresponse adjustment, by selected school-level categorical variables: 2003

School characteristic	Sample students: Eligible (percent) (*N*=6,803)	Sample students: Participating (percent) (*N*=5,456)	Bias	Relative bias	Chi-square *p*-value
School control					0.176
Public	92.77	93.78	1.01	0.011	
Private	7.23	6.22	-1.01	-0.140	
Community type					0.107
Central city	28.08	29.83	1.75	0.062	
Urban fringe/large town	38.74	39.42	0.68	0.018	
Rural/small town	33.18	30.75	-2.43	-0.073	
NAEP region					0.008
Northeast	21.02	21.21	0.19	0.009	
Southeast	29.26	25.01	-4.25	-0.145	
Central	21.78	22.94	1.16	0.053	
West	27.94	30.85	2.91	0.104	
Poverty level					0.562
High	11.25	11.66	0.41	0.036	
Low	88.75	88.34	-0.41	-0.005	

NOTE: Detail may not sum to totals because of rounding. National Assessment of Educational Progress (NAEP) region is the state-based region of the country (see appendix for state listing.) For public schools, a high poverty school is defined as one in which 50 percent or more of the students are eligible for participation in the National School Lunch Program; all private schools are treated as low poverty schools. Eligible students were 15 years old as defined by the Program for International Student Assessment (PISA). Participating students were those assessed. The relative bias is calculated as the bias divided by the estimate from the eligible sample. Students were weighted by their student nonresponse adjusted weights.
SOURCE: Organization for Economic Cooperation and Development (OECD), Program for International Student Assessment (PISA), 2003.

Based on these analyses, the chi-square statistic for NAEP region suggests that there is evidence of a relationship with participating in the assessment. There are no statistically significant relationships between participation status and school control, community type or poverty level. The bias on all four characteristics changes sign when compared to table 3-1 suggesting that the student nonresponse adjustment over-adjusted for these characteristics.

The mean number of age-eligible students for participating and eligible students is shown in table 3-7. There is no statistically significant relationship between participation status and the mean number of age-eligible students.

Table 3-7. Mean number of age-eligible students in the U.S. PISA student sample after nonresponse adjustment: 2003

	Sample students				
Student enrollment	Eligible (mean) (*N*=6,803)	Participating (mean) (*N*=5,456)	Bias	Relative bias	*t* test *p*-value
Age-eligible	329.47	338.66	9.19	0.028	0.063

NOTE: Detail may not sum to totals because of rounding**.** Eligible students were 15 years old as defined by the Program for International Student Assessment (PISA). Participating students were those assessed. The relative bias is calculated as the bias divided by the estimate from the eligible sample. Students were weighted by their student nonresponse adjusted weights.
SOURCE: Organization for Economic Cooperation and Development (OECD), Program for International Student Assessment (PISA), 2003.

Student-level Variables

The distribution of participating and eligible students by the two characteristics is shown in table 3-8. Eleven students were missing values for grade level, and two students were missing values for sex and were not included in the analysis. There are no statistically significant relationships between participation status and either of the characteristics shown in table 3-8.

Table 3-8. Percentage distribution of eligible and participating students in the U.S. PISA student sample after nonresponse adjustment, by selected student-level categorical variables: 2003

Student characteristic	Sample students: Eligible (percent)	Sample students: Participating (percent)	Bias	Relative bias	Chi-square *p*-value
Grade[1,2]					0.587
7th and 8th	3.02	2.68	-0.34	-0.113	
9th	29.83	29.71	-0.12	-0.004	
10th	60.44	60.63	0.19	0.003	
11th	6.72	6.98	0.26	0.039	
Sex[3,4]					0.565
Female	49.36	49.59	0.23	0.005	
Male	50.64	50.41	-0.23	-0.005	

[1] $N = 6{,}792$ for eligible
[2] $N = 5{,}454$ for participating
[3] $N = 6{,}801$ for eligible
[4] $N = 5{,}455$ for participating
NOTE: Information on grade level is missing for 11 of the 6,803 eligible students. Information on sex is missing for 2 of the 6,803 eligible students. Eligible students were 15-years-old as defined by the Program for International Student Assessment (PISA). Participating students were those assessed. The relative bias is calculated as the bias divided by the estimate from the eligible sample. Students were weighted by their student nonresponse adjusted weights.
SOURCE: Organization for Economic Cooperation and Development (OECD), Program for International Student Assessment (PISA), 2003.

The mean age of students (in months) for participating and eligible students is shown in table 3-9. Age was missing for one student and was excluded from the analysis. There is no statistically significant relationship between participation status and mean age of students.

Table 3-9. Mean age of eligible and participating students in the U.S. PISA student sample after nonresponse adjustment: 2003

Student age (months)	Sample students: Eligible (mean) (*N*=6,802)	Sample students: Participating (mean) (*N*=5,456)	Bias	Relative bias	*t* test *p*-value
Age	190.05	190.00	-0.05	#	0.085

Rounds to zero.
NOTE: Information on age is missing for one of the 6,803 eligible students. Eligible students were 15 years old as defined by the Program for International Student Assessment (PISA). Participating students were those assessed. The relative bias is calculated as the bias divided by the estimate from the eligible sample. Students were weighted by their student nonresponse adjusted weights.
SOURCE: Organization for Economic Cooperation and Development (OECD), Program for International Student Assessment (PISA), 2003.

Logistic Regression Model

The analysis of the student sample was intended to include a logistic regression model as was done in the school chapter with nonresponse adjustments applied. This is not possible because the number of parameters (student nonresponse adjustment classes) exceeds the number of degrees of freedom (80). The international weighting procedures created a student nonresponse adjustment class for each high/low grade combination within school. This resulted in 316 classes after collapsing. To account for the school segment of the nonresponse classes, the school characteristics that were not significant after nonresponse adjustments were applied were included in the model. To examine the relationship of the student nonresponse class to student participation, the analysis utilized a logistic regression model, with participation status as the binary dependent variable and school control, community type, poverty level, mean number of age-eligible students, and high/low grade combination as predictor variables. An *F* test was used to determine whether the student characteristics, in aggregate, were significantly related to response once the school characteristics were taken into account. The *F* test statistic is 2.19 with a *p*-value of 0.064, which indicates no significant relationship with participation.

Summary

The analysis of the PISA student nonresponse bias, using base weights, shows a statistically significant relationship between student response status and some of the available school and student characteristics. Specifically, school control, NAEP region, number of age-eligible students, sex, and the average age of students were found to be significantly different between the participating and the total eligible sample of students both in the bivariate analysis and in the multivariate logistic regression model. Further, the multivariate logistic regression model showed that students in private schools and smaller schools were more likely to participate, students in the Northeast and Central NAEP regions were less likely to participate than other students, females were more likely to participate than males, and younger students were more likely to participate than other students. Additionally, there was a significant difference found in community type in the bivariate analysis.

The bivariate analysis of the PISA student nonresponse bias, using nonresponse adjusted weights, showed significant differences between participating and the total eligible sample of students in NAEP region only. The application of student nonresponse adjustments in PISA had some effect on reducing the bias in the student characteristics and to a lesser degree in the school characteristics. None of these differences appear to be substantially large, however.

4. Session Effects

Background

Over the past decade or so, national data collection efforts that solicit the voluntary participation of schools have faced increasing resistance from school districts and schools. During this period, the Program for International Student Assessment (PISA) collected data in years 2000 and 2003 from national samples of 15-year-olds. In 2000, about 56 percent of the original sample of schools participated. Subsequent contacts with replacement schools raised this figure to 70 percent. In 2003, the response rate for the original sample dropped to 47 percent. Replacement schools were even less cooperative, as the extra effort entailed in recruiting the original schools meant that replacement schools were approached very late in the school year. However, close to 20 percent of refusals in the original sample cited time of year as the main problem and, when asked, agreed to undertake the assessment in the fall.

This phenomenon was relatively easy to explain. The past 3 years or so have seen a rapid growth in mandatory Federal and state testing as assessments have become the focal point of measures of school productivity, with specific sanctions imposed on schools failing to meet the externally established productivity levels. As a result, Federal and state assessments now have first call on the finite amount of time that schools are willing to make available for such activities in the spring of the year. States, for example, assess all students in grades 3 through 8 each year, for the specific purpose of meeting demands for increased accountability. In short, in the spring of each year, schools across the nation are focused on preparing their students for state assessments that have an important bearing on their future.

A Fall Assessment

The only viable possibility for achieving acceptable response rates under these conditions, was to conduct makeup assessments in the fall for those schools in the original sample who had indicated a willingness to participate. The age span defining eligible students for the fall sample was adjusted to ensure that, at the time of testing, they fell within the same age range as those students in the spring sample. The intention, then, was to combine the data from both spring and fall assessments, if it could be shown that there was no significant between-session bias in the achievement results of the fall sample. In the normal course of events PISA assesses 15-year-olds and, in 2003, defined 15-year-olds as those born in the 1987 calendar year. This definition was applied to the spring sample. In the case of the fall sample, this age span was adjusted such that students born between June 1987 and May 1988 were defined as eligible.

Methodology

Since schools rather than students had opted to be assessed in either the spring or the fall, there was variation among schools in this respect but not among students within schools. Thus, session effects were seen as school-level effects with the potential to influence the assessment scores of all students within a school equally but varying between schools according to whether the school was in the spring or fall session. This situation was modeled statistically as a two-level hierarchical linear model of the kind often identified as *means-as-outcomes regression* (Raudenbush and Bryk 2002, p. 24). In such models, the student-level model refers to student scores within a school that were seen as "a function of the school mean", plus a random component. In the school-level model, these school means were seen to vary between schools as a function of variation in school characteristics, *session* among them, and a random component. To the extent that session was significantly associated with variation in school means after adjustment for other (confounding) attributes of schools, a case can be made for bias in the data stemming from the fact of spring and fall assessments. Nonsignificant session effects suggest that any bias was probably inconsequential and that spring and fall data can be combined and analyzed as a whole.

Accordingly, the student-level model was fully unconditional with student scores (a function of the school mean) plus a random component. In the school-level model session, a dichotomous variable capturing *session* as spring/fall, along with a number of predictors used in the school response-bias analyses reported above, were used to predict variation in school means. The variables in question are described in the following section. The analyses were based on data from 274 schools. These schools were weighted by the school base weight.

Variables

Mathematics achievement. Since Balanced Incomplete Block (BIB)-spiraling of the assessment booklets was used to maximize subject-matter coverage, student achievement measures take the form of five plausible values for each student. For each student these were generated as random draws from an estimated ability distribution of students with similar item response patterns and backgrounds.

Session. This was a dichotomous variable indicating the assessment session in which the school participated (spring = 1 and fall = 0).

School control. This was a measure of whether the school was publicly or privately funded (public = 0, private = 1).

Number of age-eligible students. The measure was a simple count of the number of age-eligible students in the school.

Percent minority. Minority was defined as the percentage of non-White students in the school.

Community type. This was a categorical variable capturing a rural-urban dimension; central city and rural were included as dummy variables with the urban category as the omitted variable.

NAEP region. This was a categorical variable capturing four geographical regions; dummy variables for Southeast, Central, and West were included with Northeast, the omitted variable.

Note: Poverty level of the school, measured as the proportion of students eligible for participation in the National School Lunch Program, is omitted from the present model. The measure is available only for public schools. The omission seems reasonable given that much of the predictive value of this measure will be subsumed by attributes of schools already in the equation, namely, school control, percent minority, and community type.

Results

The model proposed was estimated with the HLM software (Raudenbush, Bryk, Cheong, et al. 2004). The results are shown in tables 4-1 and 4-2 in the output format used by HLM.

Table 4-1. Estimates of fixed effects for session effect model, final sample: 2003

Fixed effects	Coefficient	Standard error	*t* ratio	Degrees of freedom	*p*-value
School-level model					
Intercept	478.178	2.454	194.852	264	<0.001
Session: Spring	-3.223	6.020	-0.535	264	0.592
School control: Private	21.778	11.150	1.953	198	0.052
Age-eligible enrollment	0.100	0.022	4.539	264	<0.001
Percent minority	-1.249	0.143	-8.730	264	<0.001
Community type: Central city	-2.421	10.118	-0.239	264	0.811
Community type: Rural	-2.524	11.942	-0.211	264	0.833
NAEP region: Southeast	-11.984	11.581	-1.035	264	0.302
NAEP region: West	-5.980	11.449	-0.522	264	0.601
NAEP region: Central	-25.207	10.071	-2.503	264	0.013

NOTE: National Assessment of Educational Progress (NAEP) region is the state-based region of the country (see appendix for state listing).
SOURCE: Organization for Economic Cooperation and Development (OECD), Program for International Student Assessment (PISA), 2003.

Table 4-2. Estimates of random effects for session effect model, final sample: 2003

Random effects	Standard deviation	Variance component	Chi-square	Degrees of freedom	*p*-value
Intercept	30.753	945.755	1238.467	264	<0.001
Level-1	77.827	6057.051	†	†	†

† Not applicable.
NOTE: National Assessment of Educational Progress (NAEP) region is the state-based region of the country (see appendix A for state listing).
SOURCE: Organization for Economic Cooperation and Development (OECD), Program for International Student Assessment (PISA), 2003.

Summary

The coefficients from four of the nine predictors reach statistical significance, suggesting, respectively, that other things equal, the mean achievement of students was higher in private schools by some 22 points on average; was higher in schools with higher numbers of age-eligible students; was lower in schools with high minority enrollments; and, relative to schools in the Northeast, was lower by some 24 points on average in schools in the Central region.

Most noticeably, the coefficient for *session* fails to reach statistical significance and, thus, offers no support for the notion of statistically significant session effects. On this basis, it was determined that combining the spring and fall data would introduce no significant achievement-related bias into the combined data or the estimates derived from these data.

5. Conclusion

School Nonresponse

The investigation into nonresponse bias at the school level for the U.S. Program for International Student Assessment (PISA) effort shows no statistically significant relationship between response status and the majority of school characteristics that were examined in the analysis.

In both the original sample of schools and in the final sample, the bivariate analysis found only the *National Assessment of Educational Progress (NAEP) region* (tables 2-1 and 2-6) to be statistically significant in regard to participation. In addition, the *percentage of Black, non-Hispanic* and *Asian or Pacific Islander students* (table 2-8) was a significant factor in participation. The regression analysis, on the other hand, did not find any of the variables to be statistically significant in regard to participation. When the school nonresponse adjustments were applied to the weights for the final sample of schools, no statistically significant relationships were found between participation and any available school characteristic in the bivariate analysis, whereas the school nonresponse adjustment class was a significant predictor of school participation in the logistic regression.

These results suggest that there is little potential for nonresponse bias in the PISA original participating sample based on the characteristics studied. The results also suggest that, while there is certainly no evidence that the use of replacement schools reduced the potential for bias, it has also not added substantially to the bias. The school nonresponse adjustment was at least partially effective in reducing the nonresponse bias, in that the analysis found no significant differences among the variables (tables 2-11 through 2-14), whereas the variable forming the school nonresponse adjustment classes was found to be significant.

Student Nonresponse

The investigation into nonresponse bias at the student level for the U.S. PISA effort shows statistically significant relationships between response status and some of the school characteristics that were examined in the analysis.

The bivariate analysis, using base weights, indicates a statistically significant relationship between student response status and the following school and student characteristics: school control (table 3-1),

community type (table 3-1), NAEP region (table 3-1), number of age-eligible students (table 3-2), sex (table 3-3), and the average age of students (table 3-4). The regression approach indicated that students in private schools and smaller schools were more likely to participate, as were younger students (table 3-5).

Using nonresponse adjusted weights, one of the six variables remained statistically significant in the bivariate analysis: NAEP region (table 3-6).

These results suggest, based on the characteristics studied, that there is some potential for nonresponse bias in the PISA student sample, using base weights. None of these differences appear to be substantially large, however, especially given the large student sample sizes, which makes the tests very sensitive. The student nonresponse adjustment had some effect on reducing the bias in school and student characteristics but also tended to over-adjust for the school characteristics.

Session Bias

The analyses show no evidence of a statistically significant session effect (tables 4-1 and 4-2).

6. References

Adams, R. (Ed.). (2005). *PISA 2003 Technical Report.* Paris: Organization for Economic Cooperation and Development.

Brick, M.J., Morganstein, D., and Valliant, R. (2000). *Analysis of Complex Sample Data Using Replication.* Retrieved January 21, 2005, from http://www.westat.com/wesvar/index.html.

Ferraro, D., and Rust, K. (2003). *U.S. 2003 PISA School Sample Final Report*. Rockville, MD: Westat.

Kali, J., and Rust, K. (2005). *Analysis of School Nonresponse Bias in the 2005 NAEP –Grades 4 and 8.* (Prepared under contract to National Center for Education Statistics). Rockville, MD: Westat.

Lemke, M., Sen, A., Pahlke, E., Partelow, L., Miller, D., Williams, T., Kastberg, D., and Jocelyn, L. (2004). *International Outcomes of Learning in Mathematics Literacy and Problem Solving: PISA 2003 Results From the U.S. Perspective.* (NCES 2005-003). Washington, DC: U.S. Department of Education, National Center for Education Statistics.

Organisation for Economic Cooperation and Development (OECD). (2004). *Learning for Tomorrow's World: First Results from PISA 2003.* Paris: OECD.

Piesse, A., and Rust, K. (2003). *U.S. 2001 PIRLS Nonresponse Bias Analysis* (NCES 2003-21). U.S. Department of Education. Washington, DC: National Center for Education Statistics Working Paper.

Rao, J.N.K., and Scott, A.J. (1981). The Analysis of Categorical Data from Complex Sample Surveys: Chi-Squared Tests for Goodness of Fit and Independence in Two-Way Tables. *Journal of the American Statistical Association, 76*: 221-230.

Rao, J.N.K., and Scott, A.J. (1984). On Chi-squared Tests for Multiway Contingency Tables with Cell Proportions Estimated from Survey Data. *Annals of Statistics, 12*: 46-60.

Rao, J.N.K., and Thomas, D.R. (2003). Analysis of Categorical Response Data from Complex Surveys: an Appraisal and Update. In R.L. Chambers and C.J. Skinner (Eds.), *Analysis of Survey Data* (pp. 85-108). West Sussex, England: John Wiley and Sons.

Raudenbush, S.W., Bryk, A., Cheong, Y.F., Congdon, R., and du Toit, M. (2004). *HLM6: Hierarchical Linear and Nonlinear Modeling*. Lincolnwood, IL: Scientific Software International, Inc.

Westat. (2002). *WesVar 4.2 User's Guide.* Rockville, MD: Westat.

Wolter, K.M. (1985). *Introduction to Variance Estimation.* New York: Springer-Verlag.

Appendix A—Technical Notes

Full details on the technical aspects of the Program for International Student Assessment (PISA) 2003 can be found in Adams (2005) and Lemke et al. (2004). The sections below provide details on those aspects of the implementation of PISA that are relevant to the analyses included in this report.

Exclusions in the PISA Sample

Exclusion guidelines allowed for 0.5 percent at the school level for approved reasons (e.g., remote regions or very small schools), and 2 percent for special education schools. Overall estimated student exclusions were to be under 5 percent. PISA's intent was to be as inclusive as possible. No accommodations were offered in the United States for PISA. A special 1-hour booklet with lower difficulty items, which was scaled with the regular PISA booklets, was used in six countries for schools that would otherwise have been excluded. Special booklets were used in Austria, Belgium, the Czech Republic, Hungary, the Netherlands, and the Slovak Republic. Within schools, exclusion decisions were made by staff members who were knowledgeable about students with Individualized Education Programs (IEPs) or students who were limited English proficient, using the following international guidelines on possible student exclusions:

- *Functionally disabled students.* These were students who were permanently physically disabled in such a way that they could not perform in the testing situation. Functionally disabled students who could respond were to be included in the testing. Any sampled student who was temporarily disabled such that he or she could not participate in the assessment was considered absent from the assessment.

- *Students with mental or emotional disabilities.* These were students who were considered in the professional opinion of the school principal or by other qualified staff members to be intellectually disabled or who had been psychologically tested as such. This included students who were emotionally or mentally unable to follow even the general instructions of the test. Students were not to be excluded solely because of poor academic performance or normal disciplinary problems.

- *Students with limited proficiency in the test language.* These were students who had received less than 1 year of instruction in the language of the test. Generally, these were students who were unable to read or speak the language of the test (English in the United States) and would be unable to overcome the language barrier in the test situation.

Quality monitors from the PISA Consortium visited schools in every country to ensure testing procedures were carried out in a consistent manner across countries.

Sampling, Data Collection, and Response Rates in the United States

The 2003 PISA school sample was drawn for the United States in November 2002. The sample design for this school sample was developed to retain some of the properties of the U.S. PISA 2000 school sample, and to follow international requirements as given in the PISA sampling manual. Unlike the 2000 PISA sample, which had a three-stage design, the U.S. sample for 2003 was a two-stage sampling process with the first stage a sample of schools, and the second stage a sample of students within schools. For PISA in 2000, the U.S. school sample had the selection of a sample of geographic primary sampling units (PSUs) as the first stage of selection. The sample was not clustered at the geographic level for PISA 2003. This change was made in an effort to reduce the design effects observed in the 2000 data and to spread the respondent burden across school districts as much as possible.

The sample design for PISA was a stratified systematic sample with sampling probabilities proportional to measures of size. The PISA sample had no explicit stratification and no oversampling of subgroups. The frame was implicitly stratified (i.e., sorted for sampling) by five categorical stratification variables: grade span of the school (schools with grade 7 or 8 as last grade, grade 9 as last grade, grades 9 to 12, grades 10 to 12, and all other schools), type of school (public or private), National Assessment of Educational Progress (NAEP) region of the country[1] (Northeast, Southeast, Central, West), type of location relative to populous areas (large central city more than 250,000, mid-size central city less than 250,000, urban fringe of large central city, urban fringe of mid-size central city, large town more than 25,000, small town 2,500 to 25,000, rural outside MSA, rural inside MSA), and minority status (above or below 15 percent). The last sort key within the implicit stratification was by estimated enrollment of 15-year-olds based on grade enrollments.

At the same time that the PISA sample was selected, replacement schools were identified following the PISA guidelines by assigning the two schools neighboring the sampled school on the frame as replacements. There were several constraints on the assignment of replacements. One sampled school was not allowed to substitute for another, and a given school could not be assigned to substitute for more than one sampled school. Furthermore, replacements were required to be in the same implicit stratum as the sampled school. If the sampled school was the first or last school in the stratum, then the second school

[1] The Northeast region consists of Connecticut, Delaware, District of Columbia, Maine, Maryland, Massachusetts, New Hampshire, New Jersey, New York, Pennsylvania, Rhode Island, and Vermont. The Central region consists of Illinois, Indiana, Iowa, Kansas, Michigan, Minnesota, Missouri, Nebraska, North Dakota, Ohio, Wisconsin, and South Dakota. The West region consists of Alaska, Arizona, California, Colorado, Hawaii, Idaho, Montana, Nevada, New Mexico, Oklahoma, Oregon, Texas, Utah, Washington, and Wyoming. The Southeast region consists of Alabama, Arkansas, Florida, Georgia, Kentucky, Louisiana, Mississippi, North Carolina, South Carolina, Tennessee, Virginia, and West Virginia.

following or preceding the sampled school was identified as the replacement. One was designated a first replacement and the other a second replacement, the first replacement was then contacted. If that school also refused to participate, the second replacement school was then contacted.

The U.S. PISA school sample consisted of 420 schools. This number was increased from the international minimum requirement of 150 to offset school nonresponse, design effects, and to provide the additional students required for a metric-imperial experiment, in which some students received items using metric measurement units and other students received the same items using imperial measurement units. The student population for the PISA 2003 in the United States was the set of all 15-year-old students in the United States.[2]

The schools were selected with probability proportionate to the school's estimated enrollment of 15-year-olds from the 2003 NAEP school frame with 2000-01 school data. The data for public schools were from the Common Core of Data (CCD), and the data for private schools were from the Private School Survey (PSS). Any school containing at least one 7th through 12th grade class as of the school year 2000-01 was included on the school sampling frame. Participating schools provided lists of 15-year-old students, and a sample of 35 students was selected within each school in an equal probability sample. The overall sample design for the United States was intended to approximate a self-weighting sample of students as much as possible, with each 15-year-old student having an equal probability of being selected.

A minimum response target of 85 percent was required for initially selected educational institutions. In instances in which the initial response rate of educational institutions was between 65 and 85 percent, an acceptable school response rate could still be achieved through the use of replacement schools. Replacement schools were to be selected at the time of sample selection.

Three school response rate zones—acceptable, intermediate, and not acceptable—were defined. *Acceptable* meant that the country's data would be included in all international comparisons. *Not Acceptable* meant that the country's data would be a candidate for not being reported in international comparisons unless considerable evidence was presented that nonresponse bias was minor. *Intermediate* meant that a decision on whether or not to include the country's data in comparisons would be made while taking into account a variety of factors, such as student response rates, quality control, closeness of the response rates to the acceptable level, etc. For the purposes of calculating response rates, schools with

[2] The definition of the population of 15-year-old students was dependent upon the time of testing. A 15-year-old student tested in April and May was born in 1987, students tested in June were born between March 1987 and February 1988, and students tested in August, September, and October were born between June 1987 and May 1988.

less than 50 percent of students participating were considered not participating and their students were excluded from the student response rates. If the student response rates within such schools were at least 25 percent, these schools and students were included in the PISA 2003 database. Schools with student response rates below 25 percent were not used in any type of analysis nor are the data for these students or schools available in the PISA 2003 database.

PISA 2003 also required a minimum participation rate of 80 percent of sampled students from original and replacement schools within each country. A student was considered to be a participant if he or she participated in the first testing session or a follow-up or makeup testing session.

The PISA data collection was fielded in April, May, and June 2003. In the United States, for a variety of reasons reported by school administrators (such as increased testing requirements at the national, state, and local levels; concerns about timing of the PISA assessment; and loss of learning time), many schools in the original sample declined to participate. As it was clear that the United States would not meet the minimum response rate standards, a second testing window was opened from September to November 2003 with the agreement of the PISA Consortium, in order to improve response rates and better accommodate school schedules. For the fall data collection, the school sample included only original schools from the sample that had refused to participate in the spring but had indicated a willingness to participate in a fall assessment. Replacement schools were not included in the fall since their participation would have had little effect on raising the final response rate. In order to achieve a comparable sample of students in spring and fall, the age definition for students tested in the fall was adjusted such that all students tested were the same age. That is, in the spring sample 15-year-olds were defined as those born in the 1987 calendar year. In the case of the fall sample this age span was adjusted such that students born between June 1987 and May 1988 were defined as eligible.

Of the 420 sampled schools, 382 were eligible (some did not have any 15-year-olds enrolled) and 179 agreed to participate in the spring of 2003. An additional 70 original schools participated in the fall assessment for a total of 249 participating original schools. The school response rate (including spring and fall assessments) before replacement was 65 percent (weighted and unweighted), placing the United

States in the *intermediate* response rate category. The weighted school response rate before replacement is given by the formula:

$$\text{weighted school response rate before replacement} = \frac{\sum_{i \in Y} W_i E_i}{\sum_{i \in (Y \cup N)} W_i E_i},$$

where *Y* denotes the set of responding original sample schools with age-eligible students, *N* denotes the set of eligible nonresponding original sample schools, *Wi* denotes the base weight for school *i*, $Wi = 1/Pi$, where *Pi* denotes the school selection probability for school *i*, and *Ei* denotes the enrollment size of age-eligible students, as indicated on the sampling frame. In addition to the 249 participating original schools, 13 replacement schools also participated in the spring for a total of 262 participating schools.

A total of 7,598 students were sampled for the assessment. Of these students, 261 were deemed ineligible because of their enrolled grades, birthdays, or other reasons and were removed from the sample. Of the eligible 7,337 sampled students, an additional 534 students were excluded because of limited proficiency in English or functional or intellectual disabilities, for a weighted exclusion rate of 7 percent.

Of the 6,803 remaining sampled students, a total of 5,456 students participated in the assessment in the United States, but 114 of these came from schools that had less than 50 percent student participation. Schools that had less than 50 percent student participation were classified as nonresponding schools, and these students (114 participating students and 187 nonparticipating students) were, therefore, excluded for the purposes of calculating student response rates. Data for 5,456 students were included in the database, but student response rates were calculated by subtracting the 114 students from the 5,456 for a total of 5,342 participating students. The denominator for the student response rate was 6,502, which consist of 7,598 sampled students minus the following students: 261 ineligible; 534 excluded; 114 responding students from nonresponding schools; and 187 nonresponding students from nonresponding schools. An overall weighted student response of 83 percent was achieved (82 percent unweighted).

Since the school response rates did not meet National Center for Education Statistics (NCES) standards and fell into the *intermediate* category according to the PISA standards, NCES requested that Westat investigate the potential for bias due to nonresponse at the school and student level, along with any bias

that may have been introduced by the need to conduct the assessments in two sessions, one in the spring and the other in the fall.

The analysis of school nonresponse was conducted in two parts, examining first the original sample of schools (spring and fall participants), and then the final sample of schools (including replacements), treating as nonrespondents those schools from whom a final response was not received. Schools with 25 to 49 percent student response rates were treated as respondents in the nonresponse bias analysis since their data are included in the PISA database. Schools with student response rates less than 25 percent were treated as nonrespondents in the analysis. The analysis of school nonresponse, thus, was based on 258 and 274 responding original and final schools, respectively, of 382 eligible schools. Students in schools with 25 to 49 percent student response rates were treated as respondents in the nonresponse bias analysis again corresponding to the data included in the PISA database. The analysis of student nonresponse was based exclusively on the sample of 6,803 eligible and 5,456 participating students.. The analysis of session effects was based on the PISA definition of 274 spring and fall responding schools (including replacements).

Response Rates

Tables A-1 and A-2 details the school and student participation rates, respectively, for the PISA target population in the United States.

Table A-1. Selected characteristics of U.S. PISA analysis school sample: 2003

Schools in original sample	Eligible schools in sample	Number of participating schools before replacement	Number of participating schools after replacement	Unweighted school participation rate before replacement (percent)	Weighted school participation rate before replacement (percent)	Unweighted school participation rate after replacement (percent)	Weighted school participation rate after replacement (percent)
420	382	249	262	65.2	64.9	68.6	68.1

SOURCE: Organization for Economic Cooperation and Development (OECD), Program for International Student Assessment (PISA), 2003.

Table A-2. Selected characteristics of U.S. PISA student sample: 2003

Students sampled for assessment	Ineligible students	Excluded students	Students in nonparticipating schools excluded from response rates	Eligible students in responding schools	Number of participating students	Unweighted student participation rate after replacement (percent)	Weighted student participation rate after replacement (percent)
7,598	261	534	301	6,502	5,342	82.2	82.7

SOURCE: Organization for Economic Cooperation and Development (OECD), Program for International Student Assessment (PISA), 2003.

Description of Variables

Frame characteristics for public schools were taken from the 2000-01 CCD and, for private schools, from the 2000-01 PSS.

Race/Ethnicity

Students' race/ethnicity was obtained through student responses to a two-part question. Students were asked first whether they were Hispanic or Latino, and then asked whether they were members of the following racial groups: American Indian/Alaska Native; Asian; Black or African American; Native Hawaiian or other Pacific Islander; or White. Multiple responses to the race classification question were allowed.

Community Type

Community type was derived from the locale variable based on how the school is situated in a particular location relative to populous areas, based on the school's address. Central city consists of a large city (a principal city of a metropolitan core-based statistical area [CBSA], with the city having a population greater than or equal to 250,000) and mid-size cities (a principal city of a metropolitan CBSA, with the city having a population less than 250,000). An urban fringe/large town consists of the urban fringe of a large city (any incorporated place, census-designated place, or nonplace territory within a metropolitan CBSA of a large city and defined as urban by the Census Bureau), the urban fringe of a mid-size city (any incorporated place, census-designated place, or nonplace territory within a CBSA of a mid-size city and defined as urban by the Census Bureau) or a large town (an incorporated place or census-designated place with a population greater than or equal to 25,000 and located outside a metropolitan CBSA or inside a micropolitan CBSA). A rural/small town consists of small town (an incorporated place or census

designated place (CDP) with population less than 25,000 and greater than or equal to 2,500 and located outside a CBSA or CSA), rural, outside CBSA (any incorporated place, census-designated place, or non-place territory not within a CBSA or CSA and defined as rural by the Census Bureau) or rural, inside CBSA (any incorporated place, census-designated place, or nonplace territory within a metropolitan CBSA and defined as rural by the Census Bureau).

NAEP Region

NAEP region is the *state-based* region of the country. Northeast consists of Connecticut, Delaware, District of Columbia, Maine, Maryland, Massachusetts, New Hampshire, New Jersey, New York, Pennsylvania, Rhode Island, and Vermont. Central consists of Illinois, Indiana, Iowa, Kansas, Michigan, Minnesota, Missouri, Nebraska, North Dakota, Ohio, South Dakota, and Wisconsin. West consists of Alaska, Arizona, California, Colorado, Hawaii, Idaho, Montana, Nevada, New Mexico, Oklahoma, Oregon, Texas, Utah, Washington, and Wyoming. Southeast consists of Alabama, Arkansas, Florida, Georgia, Kentucky, Louisiana, Mississippi, North Carolina, South Carolina, Tennessee, Virginia, and West Virginia.

Statistical Procedures

Weighting

Records from the sample schools and students were assigned sampling weights to adjust for over-representation or under-representation from a particular group. The use of sampling weights is necessary for the computation of statistically sound, nationally representative estimators. The weight assigned to a school's (or student's) data is the inverse of the probability that the school (or student) would be selected for the sample. When data are weighted, each sample unit contributes to the results in proportion to the total number of schools or students represented by the individual unit. Weighting can also be used to adjust for various situations such as school and student nonresponse because data cannot be assumed to be randomly missing. The internationally defined weighting specifications for PISA require that each assessed student's sampling weight should be the product of (1) the reciprocal of the school's probability of selection, (2) an adjustment for school-level nonresponse, (3) the reciprocal of the student's probability of selection, and (4) an adjustment for student-level nonresponse.

In the analyses in this report, sometimes the appropriate weight (base weight) includes only the components of the reciprocals of the respective selection probabilities. This is the case when estimates are made based on the entire sample. In other cases nonresponse adjustments, as computed by the PISA International Study Center, are also applied. In each case the text and tables make clear which of these weighting procedures has been applied. Whereas for substantive analyses using the PISA data, one would normally apply the nonresponse adjustments when analyzing the data from the respondents in the sample, this is not always the case when carrying out analyses of potential nonresponse bias.

Sampling Errors

Sampling errors occur when the discrepancy between a population characteristic and the sample estimate arises because not all members of the reference population are sampled for the survey. The size of the sample relative to the population and the variability of the population characteristics both influence the magnitude of sampling error. The particular sample of 15-year-old students from the 2002-03 school year was just one of many possible samples that could have been selected. Therefore, estimates produced from the PISA sample may differ from estimates that would have been produced had another student sample been drawn. This type of variability is called sampling error because it arises from using a sample of 15-year-old students, rather than all 15-year-old students in that year.

The standard error is a measure of the variability due to sampling when estimating a statistic and is often included in reports containing estimates from survey data. The approach used for calculating sampling variances in PISA was the Balanced Repeated Replication (BRR)—the Fay method of BRR. In this report estimates of standard errors for each estimate are not shown. Rather the effects of sampling error are reflected in the test statistics (for *t* tests and chi-square tests, and *t* test used in logistic regression analyses) that are presented for each analysis. These are described below.

The first step to compute the variance with replication is to calculate the estimate of interest from the full sample as well as each subsample or replicate. The variation between the replicate estimates and the full-sample estimate is then used to estimate the variance for the full sample. Suppose that $\hat{\theta}$ is the full-sample estimate of some population parameter θ. The variance estimator, $v\left(\hat{\theta}\right)$, takes the form

$$v\left(\hat{\theta}\right) = c\sum_{g=1}^{G}\left(\hat{\theta}_{(g)} - \hat{\theta}\right)^2,$$

where

$\hat{\theta}_{(g)}$ is the estimate of θ based on the observations included in the g^{th} replicate,

G is the total number of replicates formed (G=80 for U.S. PISA), and

c is the constant appropriate to the replication method (c=0.05 for U.S. PISA).

The standard error is then

$$se\left(\hat{\theta}\right) = \sqrt{v\left(\hat{\theta}\right)}.$$

The Fay method of BRR algorithm used in PISA 2003 assumes that there are G replicates, each containing two sampled schools selected independently. The element $\hat{\theta}_{(g)}$ denotes the estimate using the g^{th} jackknife replicate. Each of the replicate weights are formed by multiplying the sampling weight for all cases associated with one of the units of the pair by 1.5, and the sampling weight for the elements associated with the other unit in the pair by 0.5. The determination as to which PSUs received inflated weights, and which received deflated weights, was carried out in a systematic fashion, based on the entries in a Hadamard matrix of order 80. A Hadamard matrix contains entries that are +1 and –1 in value, and has the property that the matrix, multiplied by its transpose, gives the identity matrix of order 80, multiplied by a factor of 80. (Examples of Hadamard matrices are given in Wolter 1985.) The computation of the Fay method of BRR variance for any estimate requires the computation of the statistic 81 times for any given country: once to obtain the estimate for the full sample, and 80 times to obtain the estimate for each of the jackknife replicates ($\hat{\theta}_{(g)}$).

Tests of Significance

Comparisons made in the text of this report have been tested for statistical significance. For example, when comparing results obtained from the full sample for a given grade, with those obtained only from the responding sample units, tests of statistical significance were used to establish whether or not the observed differences are statistically significant. The estimation of the standard errors that are required in order to undertake the tests of significance is complicated by the complex sample and assessment designs which both generate error variance. Together they mandate a set of statistically complex procedures in order to estimate the correct standard errors. As a consequence, the estimated standard errors contain a

sampling variance component estimated the Fay method of BRR. Details on the procedures used can be found in the *WesVar 4.0 User's Guide* (Westat 2000).

Two kinds of statistical tests are included in the report: *t* tests and chi-square tests. In addition, logistic regression analyses were conducted.

t Tests

t tests were used for testing for the hypothesis that no difference exists between the means of continuous variables for two groups (namely, the full sample and the responding sample). Suppose that $\bar{x}_A$ and $\bar{x}_B$ are the means for two groups that are being compared and $se(\bar{x}_A - \bar{x}_B)$ is the standard error of the difference between the means which accounts for the complex survey design. Then the *t* test is defined as

$$t = \frac{|\bar{x}_A - \bar{x}_B|}{se(\bar{x}_A - \bar{x}_B)}$$

This statistic is then compared to the critical values of the appropriate Student *t* distribution, to determine whether the difference is statistically significant. The appropriate number of degrees of freedom for the distribution is given by the number of primary sampling units in the design (in this case the number of schools), minus the number of sampling strata.

Note that this procedure took account of the fact that the two samples in question were not independent samples, but in fact the responding sample was a subsample of the full sample. This effect was accounted for in calculating the standard error of the difference. Note also that, in those cases where both samples were weighted just using base weights the test is exactly equivalent to testing that the mean of the respondents was equal to the mean of the nonrespondents.

Consider for example the data in table 2-2. The first row shows that the weighted mean total school enrollment for the full eligible sample of schools is 1,399.2. For the subsample of schools that participated the corresponding mean is 1,372.1, and difference of 27.0. The standard error of this estimated difference, calculated so as to reflect the dependency between these two samples, and the complex sample design, is 31.1. This gives rise to a *t* statistic of -0.87, and using 80 degrees of freedom (the appropriate figure for the PISA design), the resulting significance (or *p*-value) is 0.388. This last figure appears in the table.

t tests were also used in the logistic regression for testing for the hypothesis for whether each estimated parameter estimate is significantly different from 0. Then the *t* test is defined as

$$t = \frac{b_k}{\sqrt{v(b_k)}},$$

where b_k is a parameter estimate and $v(b_k)$ is the replication variance estimate for that parameter. This statistic is then compared to the critical values of the appropriate Student *t* distribution, as described above, to determine whether the difference is statistically significant. The appropriate number of degrees of freedom for the distribution is again given by the number of primary sampling units in the design (in this case the number of schools), minus the number of sampling strata.

Chi-square Tests

Chi-square tests are used for testing whether two distributions of a given categorical variable are different, conducted in a way that reflects the impact of the complex sample design on sampling variance. In this instance one distribution is for the full sample, and one for the responding sample. Suppose that the categorical variable in question has *c* levels, cross-tabulated producing weighted proportions *p*. The usual Pearson chi-square statistic is calculated as

$$X^2 = n\sum_{i=1}^{2}\sum_{j=1}^{c}\left(p_{ij} - p_{i.}p_{.j}\right)^2 \Big/ p_{i.}p_{.j},$$

where *j* denotes the categories of the categorical variable, and *i* indexes the samples (full sample and respondents), and *n* indicates the overall sample size. This statistic is not suitable for use directly in a statistical test with these data, for two reasons. First, the fact that the respondents are a subset of the full sample violates the standard assumptions for a chi-square test of this kind. Second, this statistic does not account for the complex sample design used to collect the data.

Thus the Pearson chi-square statistic is modified appropriately to account for the impact of these two features. The resulting test statistic is referred to as the Rao-Scott Adjusted chi-square statistic. It is sometimes also referred to as the Satterthwaite-adjusted chi-square statistic. The number of degrees of freedom for the chi-square test, normally given as (*c* -1), where *c* is the number of categories of the

categorical variable for each distribution, is also modified on account of the complex design. The modified test statistic is then compared to the chi-square distribution with the appropriate number of degrees of freedom, to determine whether the difference in the two distributions is statistically significant. For a detailed description of the technique, see Rao and Scott (1984) or Rao and Thomas (2003).

The first step in the calculation of the Satterthwaite-adjusted chi-square statistic is to form the following vector:

$$\mathrm{Y} = \sqrt{n}\begin{pmatrix} p_{11} - p_{1\cdot}p_{\cdot 1} \\ p_{12} - p_{1\cdot}p_{\cdot 2} \\ \\ p_{rc} - p_{r\cdot}p_{\cdot c} \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \\ \\ y_{rc} \end{pmatrix}$$

An *rc* x 1 vector made up of the products of the marginal proportions is defined as

$$\mathrm{p} = \begin{pmatrix} p_{1\cdot}p_{\cdot 1} \\ p_{1\cdot}p_{\cdot 2} \\ \\ p_{r\cdot}p_{\cdot c} \end{pmatrix} = \begin{pmatrix} p_1 \\ p_2 \\ \\ p_{rc} \end{pmatrix}$$

For each replicate, an *rc* x *rc* matrix is calculated whose *ij*-th element is made up of

$$\left(y_{ig} - y_i\right)\left(y_{jg} - y_j\right),$$

where y_{ig} and y_{jg} are the *i*-th and *j*-th elements of **Y** calculated for the *g*-th replicate and y_i and y_j are the corresponding full-sample values. The *ij*-th element of the estimated covariance matrix for **Y**, **B**=cov(**Y**), is calculated using the following formula:

$$B_{ij} = c\sum_{g=1}^{G}\left(y_{ig} - y_i\right)\left(y_{jg} - y_j\right),$$

where c is the constant appropriate to the replication method (c=0.05 for U.S. PISA). The Satterthwaite's approximation to degrees of freedom for the chi-square statistic to be calculated is

$$\nu = \frac{\left(\sum_{i=1}^{rc} \frac{B_{ii}}{p_i}\right)^2}{\sum_{i=1}^{rc}\sum_{j=1}^{rc} \frac{B_{ij}^2}{p_i p_j}}.$$

Since ν will generally not be an integer, interpolation in standard chi-square tables is required.

Finally, the adjusted chi-square statistic is defined as

$$RS3 = \frac{X^2}{\sum_{i=1}^{rc} \frac{B_{ii}}{p_i}}.$$

Logistic Regression Models

A linear model for investigating the relationship between binary (dichotomous) outcomes and a set of explanatory variables is referred to as a *logistic regression model*. The data are assumed to follow a binomial distribution, with probabilities that depend on the independent variables. In this instance the binary outcome of interest is whether or not the sampled unit participated in PISA.

Let p_i denote the probability that the ith sampled school will participate. Under the logistic regression model, the log odds of response propensity (expressed in terms of the logarithm of $p_i/(1-p_i)$), is assumed to have the following linear form:

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_p X_{pi}$$

where X_{1i}, X_{2i} ..., X_{pi} are p auxiliary variables associated with the ith sampled beneficiary, and β_0, β_1, ..., β_p are coefficients to be estimated. Asymptotic assumptions are used to develop statistical tests to determine which, if any, of the coefficients are significantly different from zero. In the analyses in

this report the standard procedures for carrying out logistic regression analyses have been modified both to incorporate the sampling weights in the estimation of the coefficients, and to reflect the effect of the complex sample design on the variance-covariance matrix of the coefficients.

The Newton-Raphson algorithm is used to iteratively solve for parameter solutions in the logistic regression. Let $q(\boldsymbol{\beta}) = \partial L_n(\boldsymbol{\beta})/\partial\boldsymbol{\beta}$ be the vector of first partial derivatives of the sample log-likelihood with respect to $\boldsymbol{\beta}$. Let $\mathbf{H}(\boldsymbol{\beta})$ be the matrix of second partial derivatives (or Hessian) of the sample log-likelihood having entries $\partial^2 \mathrm{L}/\partial\beta_a\partial\beta_b$, where β_a and β_b are two separate components of $\boldsymbol{\beta}$. Denote by $\mathbf{q}^t$ and $\mathbf{H}^t$ the values of $q(\boldsymbol{\beta})$ and $\mathbf{H}(\boldsymbol{\beta})$ evaluated at $\mathbf{b}^t$, the value of the estimate $\mathbf{b}$ at step t.

The general approach is to approximate the sample log-likelihood at the desired estimate, $L_n(\mathbf{b})$, at step t in the iterative process near the point $\mathbf{b}^t$ by a second-order Taylor series expansion:

$$L_n^t(\mathbf{b}) \cong L_n(\mathbf{b}^t) + \mathbf{q}^{t\prime}(\mathbf{b}-\mathbf{b}^t) + \tfrac{1}{2}(\mathbf{b}-\mathbf{b}^t)'\,\mathbf{H}^t(\mathbf{b}-\mathbf{b}^t).$$

Solving $\partial L^t/\partial\mathbf{b} = \mathbf{q}^t + \mathbf{H}^t(\mathbf{b}-\mathbf{b}^t) = \mathbf{0}$ for $\mathbf{b}$ yields the iteration equations

$$\mathbf{b}^{t+1} = \mathbf{b}^t - \left[\mathbf{H}^t\right]^{-1}\mathbf{q}^t,$$

assuming $\mathbf{H}^t$ has an inverse. Given an initial value for $t = 0$, the set of iteration equations is solved for $\mathbf{b}^1$, $\mathbf{b}^1$ is used to solve for $\mathbf{b}^2$, and so on, until the convergence criterion is satisfied. The $se(\hat{\beta})$ is calculated using the Fay method of BRR and repeating the procedure for each replicate.

Hierarchical Linear Models

Hierarchical Linear Models (HLM) is a class of techniques for analyzing data that have a hierarchical or nested structure. Typically in education data students are nested within schools and schools within districts. (There can also be further nesting for example of observations of multiple time points within each child or of districts within PSUs.)

HLM is appropriate for this sort of analysis for major two reasons. First, observations nested within groups are correlated due to the common context shared by students within the same group. HLM is

designed to provide proper estimates of standard errors that account for the clustering of outcomes within groups. Second, HLM allows for the modeling of responses at each level of the hierarchy. To illustrate, consider a two-level model with students nested within schools. At the first level one is interested in estimating the relationship between how many math courses are taken in high school and math achievement:

Level 1- Students nested within schools:

$$yij = \beta 0j + \beta 1jX1ij + \beta 2jX2ij + ... + \beta pjXpij + eij$$

where,

y_{ij} is the achievement of child *i* in school *j*,
β_{0j} is the school regression intercept,
β_{1j} is the regression coefficient of the relationship between X_1 and achievement (controlled for X_2),
β_2 is the regression coefficient of X_2 predicting achievement, and
e_{ij} is a random error.

The distinctive feature of HLM is that the first level parameters can be modeled in a group-level regression. At level 2, adjusted school means, β_{0j}, and regression slopes, β_{1j} are predicted by school factors.

Level 2- Schools:

$$\beta 0j = \gamma 00 + \gamma 01Z1j + \gamma 02Z2j + ... + \gamma 0qZqj + u0j$$
$$\beta 1j = \gamma 10 + u1j$$

Where γ's are school-level regression coefficients for *Z* factors and the *u*'s are random errors.