Appendix A: Methodology and Technical Notes

This appendix describes features of the Program for International Student Assessment (PISA) 2012 methodology, including sample design, test design, and scoring, with a particular focus on the U.S. implementation. For further details about the assessment and any of the topics discussed here, see the Organization for Economic Cooperation and Development's (OECD) *PISA 2012 Technical Report* (forthcoming).

International Requirements for Sampling, Data Collection, and Response Rates

OECD required all participating education systems to adhere to the PISA 2012 technical standards (OECD forthcoming), which provided detailed information about the target population, sampling, response rates, translation, assessment administration, and data submission. According to the standards, the international desired population in each education system consisted of 15-year-olds attending both publicly and privately controlled schools in grade 7 and higher. To provide valid estimates of student achievement and characteristics, the sample of PISA students had to be selected in a way that represented the full population of 15-year-old students in each education system. The sample design for PISA 2012 was a stratified systematic sample, with sampling probabilities proportional to the estimated number of 15-year-old students in the school based on grade enrollments. Samples were drawn using a two-stage sampling process. The first stage was a sample of schools, and the second stage was a sample of students within schools. The PISA international contractors responsible for the design and implementation of PISA internationally (hereafter referred to as the PISA consortium) drew the sample of schools for each education system.

A minimum of 4,500 students from a minimum of 150 schools was required in each country.[1] Following the PISA consortium guidelines, replacement schools were identified at the same time the PISA sample was selected by assigning the two schools neighboring the sampled school in the frame as replacements. The international guidelines specified that within schools, a sample of 35 students was to be selected in an equal probability sample unless fewer than 35 students age 15 were available (in which case all 15-year-old students were selected).

Each education system collected its own data, following international guidelines and specifications. The technical standards required that students in the sample be 15 years and 3 months to 16 years and 2 months at the beginning of the testing period. The maximum length of the testing period was 42 days. Most education systems conducted testing from March through August 2012.[2]

The school response-rate target was 85 percent for all education systems. This target applies in aggregate, not to each individual school. A minimum of 65 percent of schools from the original sample of schools was required to participate for an education system's data to be included in the international database. Education systems were allowed to use replacement

[1] PISA also includes education systems that are not countries, such as Hong Kong and Shanghai in China. Non-national entities were required to sample a minimum of 1,500 students from at least 50 schools. In the United States, three states (Connecticut, Florida, and Massachusetts) provided state-level samples in addition to the schools for the national sample in order to obtain state-level PISA estimates.

[2] The United States and the United Kingdom were given permission to move the testing dates to September through November in an effort to improve response rates. The range of eligible birth dates was adjusted so that the mean age remained the same (i.e., 15 years and 3 months to 16 years and 2 months at the beginning of the testing period). In 2003, the United States conducted PISA in the spring and fall and found no significant difference in student performance between the two time points.

schools (selected during the sampling process) to increase the response rate once the 65 percent benchmark had been reached. Replacement students within a school were not allowed.

The technical standards also required a minimum participation rate of 80 percent of sampled students from schools (sampled and replacement) within each education system. Follow-up sessions were required in schools where too few students participated in the originally scheduled test sessions to ensure a high overall student response rate. 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. Data from education systems not meeting this requirement could be excluded from international reports. See appendix B for final response rates by education system.

PISA 2012 is designed to be as inclusive as possible. The guidelines allowed schools to be excluded for approved reasons (for example, schools in remote regions, very small schools, or special education schools). Schools used the following international guidelines on student exclusions:

- **Students with functional disabilities.** These were students with a moderate to severe permanent physical disability such that they cannot perform in the PISA testing environment.
- **Students with intellectual disabilities.** These were students with a mental or emotional disability and who have been tested as cognitively delayed or who are considered in the professional opinion of qualified staff to be cognitively delayed such that they cannot perform in the PISA testing environment.
- **Students with insufficient language experience.** These were students who meet the three criteria of not being native speakers in the assessment language, having limited proficiency in the assessment language, and having less than 1 year of instruction in the assessment language.

Overall estimated exclusions (including both school and student exclusions) were to be under 5 percent of the PISA target population.

Sampling and Data Collection in the United States

The PISA 2012 school sample was drawn for the United States by the PISA consortium. The U.S. PISA sample was stratified into eight explicit groups based on control of school (public or private) and region of the country (Northeast, Central, West, Southeast).[3] Within each stratum, the frame was sorted for sampling by five categorical stratification variables: grade range of the school (five categories); type of location relative to populous areas (city, suburb, town, rural);[4] combined percentage of Black, Hispanic, Asian, Native Hawaiian/Pacific Islander, and American Indian/Alaska Native students (above or below 15 percent); gender (mostly female [percent female >= 95 percent], mostly male [percent female < 5 percent]; and other); and state. The same frame and characteristics were used for the state samples.

For the U.S. national sample, within each school, 50 students aged 15 were randomly sampled. The United States increased its national sample from the international standard of 35 to 50 in order

[3] The Northeast region consists of Connecticut, Delaware, the 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.

[4] These types are defined as follows: (1) "city" is a territory inside an urbanized area with a core population of 50,000 or more and inside a principal city; (2) "suburb" is a territory inside an urbanized area with a core population of 50,000 or more and outside a principal city; (3) "town" is a territory inside an urban cluster with a core population between 25,000 and 50,000; and (4) "rural" is a territory not in an urbanized area or urban cluster.

to reach the required number of students and in order to administer the optional financial literacy assessment. Connecticut, Florida, and Massachusetts participated in PISA 2012 with separate state samples drawn by the PISA consortium. The state samples are not part of the main sample. In each of the three state samples, 42 students aged 15 were randomly sampled within each school. If fewer than 50 age-eligible students (in schools in the national sample) or fewer than 42 age-eligible students (in schools in the state samples) were enrolled, all 15-year-old students in a school were selected. Thus, in each school, each age-eligible student had an equal probability of being selected. Sampled students were born between July 1, 1996, and June 30, 1997 (hereafter the sampled students are referred to as "15-year-olds" or "15-year-old students").

In the national-sample schools, of the 50 students, 42 took the paper-based mathematics, science, and reading literacy assessments, 20 of which were subsampled to also take the computer-based assessment, and 8 took the financial literacy assessment. In the state sample schools, all sampled students took only the paper-based mathematics, science, and reading assessments. The technical standard for the maximum length of the testing period was 42 days, but the United States requested and was granted permission to expand the testing window to 60 days (from October 2, 2012, to November 30, 2012) to accommodate school requests.

The U.S. PISA 2012 national school sample consisted of 240 schools.[5] This number was increased from the international minimum requirement of 150 to offset school nonresponse and reduce design effects. Schools were selected with probability proportionate to the school's estimated enrollment of 15-year-olds. The data for public schools were from the 2008–09 Common Core of Data and the data for private schools were from the 2009–10 Private School Universe Survey. Any school containing at least one 7th- through 12th-grade class was included in the school sampling frame. Participating schools provided a list of 15-year-old students (typically in August or September 2012) from which the sample was drawn using sampling software provided by the international contractor.

In addition to the international response rate standards described in the prior section, the U.S. sample had to meet the statistical standards of the National Center for Education Statistics (NCES) of the U.S. Department of Education. For assessments, NCES requires that the response rate should be at least 80 percent for schools and at least 85 percent for students.

Test Development

The 2012 assessment instruments were developed by international experts and PISA consortium test developers and included items submitted by participating education systems. Items were reviewed by representatives of each country for possible bias and relevance to PISA's goals and the PISA subject-matter expert groups. All participating education systems field-tested the assessment items in spring 2011.

The final paper-based assessment consisted of 85 mathematics items, 44 reading items, 53 science items, and 40 financial literacy items allocated to 17 test booklets (in education systems that did not administer the optional financial literacy assessment there were 13 test booklets). Each booklet was made up of four test clusters. Altogether there were seven mathematics clusters, three reading clusters, three science clusters, and two financial literacy clusters. The mathematics, science, and reading clusters were allocated in a rotated design to 13 booklets. The financial literacy clusters

[5] The state samples consisted of 54, 55, and 54 schools for Connecticut, Florida, and Massachusetts, respectively. As with the PISA national sample, these numbers were increased from the international minimum of 50 schools for subnational entities to offset school nonresponse and ineligibility.

in conjunction with mathematics and reading clusters were allocated in a rotated design to four booklets. The average number of items per cluster was 12 items for mathematics, 15 items for reading, 18 items for science, and 20 items for financial literacy. Each cluster was designed to average 30 minutes of test material. Each student took one booklet, with about 2 hours' worth of testing material. Approximately half of the items were multiple-choice, about 20 percent were closed or short response types (for which students wrote an answer that was simply either correct or incorrect), and about 30 percent were open constructed responses (for which students wrote answers that were graded by trained scorers using an international scoring guide). In PISA 2012, with the exception of students participating in the financial literacy assessment, every student answered mathematics items. Not all students answered reading, science items, and\or financial literacy items.

A subset of students who took the paper-based assessment also took a 40-minute computer-based assessment. In the United States, the computer-based assessment consisted of problem solving and the optional computer-based assessment of mathematics and reading. The computer-based assessment consisted of 168 problem-solving items, 164 mathematics items, and 144 reading items allocated to 24 forms. Each form was made up of two clusters that together contained 18 to 22 items. Altogether there were four clusters of problem solving, four clusters of mathematics, and two clusters of reading. The problem-solving, mathematics, and reading clusters were allocated in a rotated design to the 24 forms. Each cluster was designed to average 20 minutes of test material. (Not all education systems participated in the computer-based assessment and some education systems only administered the computer-based problem-solving assessment. Education systems that administered only the problem-solving assessment followed a different rotation design.)

In addition to the cognitive assessment, students also completed a 30-minute questionnaire designed to provide information about their backgrounds, attitudes, and experiences in school. Principals in schools where PISA was administered also completed a 30-minute questionnaire about their schools.

Translation and Adaptation

Source versions of all instruments (assessment booklets, computer-based assessment forms, questionnaires, and manuals) were prepared in English and French and translated into the primary language or languages of instruction in each education system. The PISA consortium recommended that education systems prepare and consolidate independent translations from both source versions and provided precise translation guidelines that included a description of the features each item was measuring and statistical analysis from the field trial. In cases for which one source language was used, independent translations were required and discrepancies reconciled. In addition, it was sometimes necessary to adapt the instrument for cultural purposes, even in nations such as the United States that use English as the primary language of instruction. For example, words such as "lift" might be adapted to "elevator" for the United States. The PISA consortium verified the national adaptation of all instruments. Electronic copies of printed materials were sent to the PISA consortium for a final visual check prior to data collection.

Test Administration and Quality Assurance

The PISA consortium emphasized the use of standardized procedures in all education systems. Each education system collected its own data, based on a manual provided by the PISA consortium (ACER 2011) that explained the survey's implementation, including precise instructions for the work of school coordinators and scripts for test administrators to use in testing sessions. Test administration

in the United States was conducted by professional staff trained in accordance with the international guidelines. Students were allowed to use calculators, and U.S. students were provided calculators.

In a sample of schools in each education system, a PISA Quality Monitor (PQM) who was engaged by the PISA consortium observed test administrations. The sample schools were selected jointly by the PISA consortium and the PQM. In the United States, there were two PQMs who each observed seven schools from the national and state samples. The PQM's primary responsibility was to document the extent to which testing procedures in schools were implemented in accordance with test administration procedures. The PQM's observations in U.S. schools indicated that international procedures for data collection were applied consistently.

Weighting

The use of sampling weights is necessary for the computation of statistically sound, nationally representative estimates. Adjusted survey weights adjust for the probabilities of selection for individual schools and students, for school or student nonresponse, and for errors in estimating the size of the school or the number of 15-year-olds in the school at the time of sampling. Survey weighting for all education systems participating in PISA 2012 was coordinated by Westat, as part of the PISA consortium.

The school base weight was defined as the reciprocal of the school's probability of selection multiplied by the number of eligible students in the school. (For replacement schools, the school base weight was set equal to the original school it replaced.) The student base weight was given as the reciprocal of the probability of selection for each selected student from within a school.

The product of these base weights was then adjusted for school and student nonresponse. The school nonresponse adjustment was done individually for each education system by cross-classifying the explicit and implicit stratification variables defined as part of the sample design. Usually about 10 to 15 such cells were formed per education system.

The student nonresponse adjustment was done within cells based first on their school nonresponse cell and their explicit stratum; within that, grade and sex were used when possible. All PISA analyses were conducted using these adjusted sampling weights. For more information on the nonresponse adjustments, see OECD's *PISA 2012 Technical Report* (forthcoming).

Scaling of Student Test Data

Each test booklet or computerized version had a different subset of items. The fact that each student completed only a subset of items means that classical test scores, such as the percentage correct, are not accurate measures of student performance. Instead, scaling techniques were used to establish a common scale for all students. For PISA 2012, item response theory (IRT) was used to estimate average scores for mathematics, science, and reading literacy for each education system, as well as for three mathematics process and four mathematics content scales. For education systems participating in the financial literacy assessment and the computer-based assessment, these assessments will be scaled separately and assigned separate scores.

IRT identifies patterns of response and uses statistical models to predict the probability of answering an item correctly as a function of the students' proficiency in answering other questions. With this method, the performance of a sample of students in a subject area or subarea can be summarized on a simple scale or series of scales, even when students are administered different items.

Scores for students are estimated as plausible values because each student completed only a subset of items. Five plausible values were estimated for each student for each scale. These values represent the distribution of potential scores for all students in the population with similar characteristics and identical patterns of item response. Statistics describing performance on the PISA reading, mathematics, and science literacy scales are based on plausible values.

Proficiency Levels

In addition to a range of scale scores as the basic form of measurement, PISA describes student proficiency in terms of levels. Higher levels represent the knowledge, skills, and capabilities needed to perform tasks of increasing complexity. PISA results are reported in terms of percentages of the student population at each of the predefined levels.

To determine the performance levels and cut scores on the literacy scales, IRT techniques were used. With IRT techniques, it is possible to simultaneously estimate the ability of all students taking the PISA assessment, as well as the difficulty of all PISA items. Estimates of student ability and item difficulty can then be mapped on a single continuum. The relative ability of students taking a particular test can be estimated by considering the percentage of test items they get correct. The relative difficulty of items in a test can be estimated by considering the percentage of students getting each item correct. In PISA, all students within a level are expected to answer at least half of the items from that level correctly. Students at the bottom of a level are able to provide the correct answers to about 52 percent of all items from that level, have a 62 percent chance of success on the easiest items from that level, and have a 42 percent chance of success on the most difficult items from that level. Students in the middle of a level have a 62 percent chance of correctly answering items of average difficulty for that level (an overall response probability of 62 percent). Students at the top of a level are able to provide the correct answers to about 70 percent of all items from that level, have a 78 percent chance of success on the easiest items from that level, and have a 62 percent chance of success on the most difficult items from that level. Students just below the top of a level would score less than 50 percent on an assessment at the next higher level. Students at a particular level demonstrate not only the knowledge and skills associated with that level but also the proficiencies defined by lower levels. Patterns of responses for students below level 1b for reading literacy and below level 1 for mathematics and science literacy suggest that these students are unable to answer at least half of the items from those levels correctly. For details about the approach to defining and describing the PISA levels and establishing the cut scores, see the OECD's *PISA 2012 Technical Report* (forthcoming). The table on the following page shows the cut scores for each proficiency level for mathematics, science, and reading literacy.

Table A-1. Cut scores for proficiency levels for mathematics, science, and reading literacy: 2012

Proficiency level	Mathematics	Science	Reading[1]
Below level 1	0-357.77	0-334.94	0-262.04
Level 1	greater than 357.77-420.07	greater than 334.94-409.54	greater than 262.04-334.75 (1b) greater than 334.75-407.47 (1a)
Level 2	greater than 420.07-482.38	greater than 409.54-484.14	greater than 407.47-480.18
Level 3	greater than 482.38-544.68	greater than 484.14-558.73	greater than 480.18-552.98
Level 4	greater than 544.68-606.99	greater than 558.73-633.33	greater than 552.98-625.61
Level 5	greater than 606.99-669.30	greater than 633.33-707.93	greater than 625.61-698.32
Level 6	greater than 669.30-1000	greater than 707.93-1000	greater than 698.32-1000

[1]The first reading literacy proficiency level is composed of levels 1a and 1b. The score range for below level 1 refers to scores below level 1b.
SOURCE: Organization for Economic Cooperation and Development (OECD), Program for International Student Assessment (PISA), 2012.

Data Limitations

As with any study, there are limitations to PISA 2012 that should be taken into consideration. Estimates produced using data from PISA 2012 are subject to two types of error: nonsampling errors and sampling errors.

Nonsampling error is a term used to describe variations in the estimates that may be caused by population coverage limitations, nonresponse bias, and measurement error, as well as data collection, processing, and reporting procedures. For example, suppose the study was unsuccessful in getting permission from many rural schools in a certain region of the country. In that case, reports of means for rural schools for that region may be biased. Fortunately, such a coverage problem did not occur in PISA in the United States. The sources of nonsampling errors are typically problems such as unit and item nonresponse, the differences in respondents' interpretations of the meaning of survey questions, and mistakes in data preparation.

Sampling errors arise when a sample of the population, rather than the whole population, is used to estimate some statistic. Different samples from the same population would likely produce somewhat different estimates of the statistic in question. This fact means that there is a degree of uncertainty associated with statistics estimated from a sample. This uncertainty is referred to as sampling variance and is usually expressed as the standard error of a statistic estimated from sample data. The approach used for calculating standard errors in PISA was the Fay method of balanced repeated replication (BRR) (Judkins 1990). This method of producing standard errors uses information about the sample design to produce more accurate standard errors than would be produced using simple random sample assumptions.

Standard errors can be used as a measure for the precision expected from a particular sample. Standard errors for all statistics reported in this report are available at http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2014024.

Confidence intervals provide a way to make inferences about population statistics in a manner that reflects the sampling error associated with the statistic. Assuming a normal distribution and a 95 percent confidence interval, the population value of this statistic can be inferred to lie within the confidence interval in 95 out of 100 replications of the measurement on different samples drawn from the same population.

Confidentiality and Disclosure Limitations

Confidentiality analyses for the United States were designed to provide reasonable assurance that public-use data files issued by the PISA consortium would not allow identification of individual U.S. schools or students when compared against other public-use data collections. Disclosure limitations included identifying and masking potential disclosure risk to PISA schools and including an additional measure of uncertainty to school and student identification through random swapping of data elements within the student and school file. Swapping was designed to not significantly affect estimates of means and variances for the whole sample or reported subgroups (Krenzke et al. 2006).

Statistical Procedures

Comparisons made in the text of this report have been tested for statistical significance. For example, in the commonly made comparison of OECD averages to U.S. averages, tests of statistical significance were used to establish whether or not the observed differences from the U.S. average were statistically significant.

In almost all instances, the tests for significance used were standard *t* tests. These fell into two categories according to the nature of the comparison being made: comparisons of independent samples and comparisons of nonindependent samples. In PISA, education system groups are independent. We judge that a difference is "significant" if the probability associated with the *t* test is less than .05. If a test is significant this implies that difference in the observed means in the sample represents a real difference in the population.[6] No adjustments were made for multiple comparisons.

In simple comparisons of independent averages, such as the average score of education system 1 with that of education system 2, the following formula was used to compute the *t* statistic:

$$t = \frac{(est_1 - est_2)}{\sqrt{se_1^2 + se_2^2}},$$

where est_1 and est_2 are the estimates being compared (e.g., averages of education system 1 and education system 2) and se_1^2 and se_2^2 are the corresponding squared standard errors of these averages. The PISA 2012 data are hierarchical and include school and student data from the participating schools. The standard errors for each education system take into account the clustered nature of the sampled data. These standard errors are not adjusted for correlations between groups since groups are independent.

The second type of comparison occurs when evaluating differences between nonindependent groups within the education system. Because of the sampling design in which schools and students within schools are randomly sampled, the data within the education system from mutually exclusive sets of students (for example, males and females) are not independent. As a result, to determine whether the performance of females differs from the performance of males, for example, the standard error of the difference taking into account the correlation between females' scores and males' scores needs to be estimated. A BRR procedure, described above, was used to estimate the standard errors of differences

[6] A .05 probability implies that the *t* statistic is among the 5 percent most extreme values one would expect if there were no difference between the means. The decision rule is that when *t* statistics are this extreme, they are sampled from a population where there is a difference between the means.

between nonindependent samples within the United States. Use of the BRR procedure implicitly accounts for the correlation between groups when calculating the standard errors.

To test comparisons between nonindependent groups the following *t* statistic formula was used:

$$t = \frac{(est_{grp1} - est_{grp2})}{se_{(grp1 - grp2)}},$$

where est_{grp1} and est_{grp2} are the nonindependent group estimates being compared and $se_{(grp1\text{-}grp2)}$ is the standard error of the difference calculated using BRR to account for the correlation between the estimates for the two nonindependent groups.

This page intentionally left blank

Appendix B: International and U.S. Data Collection Results

This appendix describes the success of participating education systems in meeting the international technical standards on data collection described in appendix A. Information is provided for all participating education systems on their coverage of the target population, exclusion rates, and response rates. This appendix also provides the U.S. response rates and the results of the U.S. nonresponse bias analysis.

Response Rates

Table B-1 provides information on weighted school participation rates before and after school replacement and the number of participating schools after replacement for each participating education system. Table B-2 provides information on coverage of the target population, overall exclusion rates, weighted student response rates after school replacement, and the number of participating students after replacement for each participating education system.

One hundred thirty-nine participating original schools and 23 replacement schools participated in the U.S. administration of the Program for International Student Assessment (PISA), for a total of 162 schools. Although all 162 schools were included in the analysis of the U.S. PISA 2012 results, international guidelines stipulated that schools with between 25 and 50 percent of students participating were considered nonparticipating schools for the purposes of calculating response rates (but were eligible to be included in the analysis of results). In the United States, one replacement school had a student response rate between 25 and 50 percent. This resulted in 161 participating schools and an overall weighted school response rate of 77 percent. The overall weighted student response rate was 89 percent and the U.S. overall student exclusion rate was 5 percent.

In Connecticut, there were 50 participating schools (out of 51 eligible schools), resulting in an overall weighted school response rate of 98 percent. The overall weighted student response rate was 87 percent and the overall student exclusion rate was 4 percent. In Florida, there were 54 participating schools (out of 54 eligible schools), resulting in an overall weighted school response rate of 100 percent. The overall weighted student response rate was 90 percent and the overall student exclusion rate was 8 percent. In Massachusetts, there were 49 participating schools (out of 49 eligible schools), resulting in an overall weighted school response rate of 100 percent. The overall weighted student response rate was 90 percent and the overall student exclusion rate was 4 percent.

Table B-1. Number of schools and weighted participation rates, by education system: 2012

Education system	Percent		Number of participating schools after replacement
	Weighted school participation before replacement	Weighted school participation after replacement	
Albania	100.0	100.0	204
Argentina	95.5	95.9	219
Australia	97.9	97.9	757
Austria	100.0	100.0	191
Belgium	84.4	96.6	282
Brazil	92.7	95.4	837
Bulgaria	99.2	99.8	187
Canada	91.3	92.9	840
Chile	91.9	98.8	221
Chinese Taipei	100.0	100.0	163
Colombia	86.6	97.4	352
Costa Rica	98.9	98.9	191
Croatia	98.7	99.9	163
Cyprus	96.6	96.6	117
Czech Republic	98.1	99.6	295
Denmark	87.0	95.5	339
Estonia	100.0	100.0	206
Finland	99.0	99.3	311
France	96.6	96.6	223
Germany	97.7	98.0	228
Greece	93.2	98.9	188
Hong Kong-China	78.7	94.1	147
Hungary	97.6	99.4	204
Iceland	99.3	99.3	133
Indonesia	94.9	98.0	206
Ireland	98.7	99.3	183
Israel	91.1	93.8	172
Italy	89.1	97.4	1,186
Japan	86.3	95.5	191
Jordan	100.0	100.0	233
Kazakhstan	100.0	100.0	218
Korea, Republic of	99.9	99.9	156
Latvia	87.9	99.9	211
Liechtenstein	100.0	100.0	12
Lithuania	98.2	100.0	216
Luxembourg	100.0	100.0	42
Macao-China	100.0	100.0	45

See notes at end of table.

Table B-1. Number of schools and weighted participation rates, by education system: 2012—Continued

Education system	Percent		Number of participating schools after replacement
	Weighted school participation before replacement	Weighted school participation after replacement	
Malaysia	100.0	100.0	164
Mexico	91.8	95.3	1,468
Montenegro, Republic of	100.0	100.0	51
Netherlands	75.3	89.4	177
New Zealand	80.9	89.3	177
Norway	85.2	94.7	197
Peru	97.9	98.6	240
Poland	85.4	97.9	182
Portugal	95.4	95.8	187
Qatar	99.9	99.9	157
Romania	100.0	100.0	178
Russian Federation	100.0	100.0	227
Serbia, Republic of	90.0	95.4	152
Shanghai-China	100.0	100.0	155
Singapore	97.5	98.2	172
Slovak Republic	87.5	99.0	231
Slovenia	98.1	98.1	335
Spain	99.7	99.7	902
Sweden	98.9	99.8	209
Switzerland	94.5	98.3	410
Thailand	98.0	100.0	239
Tunisia	99.3	99.3	152
Turkey	97.5	99.9	169
United Arab Emirates	99.4	99.4	453
United Kingdom	80.1	89.2	505
United States	**67.1**	**77.2**	**161**
Uruguay	99.4	100.0	180
Vietnam	100.0	100.0	162
U.S. state education systems			
Connecticut	98.0	98.0	50
Florida	100.0	100.0	54
Massachusetts	100.0	100.0	49

NOTE: In calculating school participation rates, each school received a weight equal to the product of its base weight (the reciprocal of its probability of selection) and the number of age-eligible students enrolled in the school, as indicated on the sampling frame. Weighted school participation before replacement refers to the sum of weights of the original sample schools with PISA-assessed students and a student response rate of at least 50 percent over the sum of weights of all original sample schools. Weighted school participation after replacement refers to the sum of weights of the original and replacement schools with PISA-assessed students and a student response rate of at least 50 percent over the sum of weights of responding original sample schools, responding replacement schools, and eligible refusing original sample schools. Italics indicate non-OECD countries and education systems. Results for Connecticut, Florida, and Massachusetts are for public schools only.
SOURCE: Organization for Economic Cooperation and Development (OECD), Program for International Student Assessment (PISA), 2012.

Table B-2. Coverage of target population, student exclusion and weighted participation rates, and number of students, by education system: 2012

Education system	Percent			Number of participating students
	Coverage of national desired population	Overall student exclusion rate	Weighted student participation after replacement	
Albania	99.9	0.1	92.5	4,743
Argentina	99.3	0.7	88.0	5,908
Australia	96.0	4.0	86.8	17,774
Austria	98.7	1.3	91.7	4,756
Belgium	98.6	1.4	90.9	9,690
Brazil	98.6	1.4	90.1	20,091
Bulgaria	97.4	2.6	95.7	5,282
Canada	93.6	6.4	80.8	21,548
Chile	98.7	1.3	94.6	6,857
Chinese Taipei	98.8	1.2	96.3	6,046
Colombia	99.9	0.1	93.1	11,173
Costa Rica	100.0	0.0	89.0	4,602
Croatia	97.8	2.2	92.2	6,153
Cyprus	96.7	3.3	93.3	5,078
Czech Republic	98.2	1.8	90.1	6,535
Denmark	93.8	6.2	89.1	7,481
Estonia	94.2	5.8	92.9	5,867
Finland	98.1	1.9	90.7	8,829
France	95.6	4.4	89.5	5,682
Germany	98.5	1.5	93.2	5,001
Greece	96.4	3.6	96.7	5,125
Hong Kong-China	98.2	1.8	93.1	4,670
Hungary	97.4	2.6	92.7	4,810
Iceland	96.2	3.8	84.7	3,508
Indonesia	99.7	0.3	95.2	5,622
Ireland	95.5	4.5	84.1	5,016
Israel	95.9	4.1	90.0	6,061
Italy	96.7	3.3	92.8	38,142
Japan	97.9	2.1	96.1	6,351
Jordan	99.6	0.4	95.0	7,038
Kazakhstan	96.6	3.4	98.9	5,808
Korea, Republic of	99.2	0.8	98.7	5,033
Latvia	96.0	4.0	90.9	5,276
Liechtenstein	95.8	4.2	93.3	293
Lithuania	96.0	4.0	92.1	4,618
Luxembourg	87.2	8.4	95.2	5,260
Macao-China	99.8	0.2	99.4	5,335
Malaysia	99.8	0.2	94.0	5,197
Mexico	99.3	0.7	93.9	33,806

See notes at end of table.

Table B-2. Coverage of target population, student exclusion and weighted participation rates, and number of students, by education system: 2012—Continued

Education system	Percent			Number of participating students
	Coverage of national desired population	Overall student exclusion rate	Weighted student participation after replacement	
Montenegro, Republic of	99.7	0.3	93.8	4,744
Netherlands	95.6	4.4	85.0	4,460
New Zealand	95.4	4.6	84.7	5,248
Norway	93.9	6.1	90.9	4,686
Peru	99.8	0.2	96.0	6,035
Poland	95.4	4.6	87.6	5,662
Portugal	98.4	1.6	87.4	5,722
Qatar	97.5	2.5	99.7	10,966
Romania	96.5	3.5	97.8	5,074
Russian Federation	97.6	2.4	97.3	6,418
Serbia, Republic of	97.1	2.9	93.4	4,684
Shanghai-China	98.5	1.5	98.5	6,374
Singapore	98.8	1.2	94.3	5,546
Slovak Republic	97.1	2.9	93.8	5,737
Slovenia	98.4	1.6	90.5	7,229
Spain	95.7	4.3	89.9	25,335
Sweden	94.6	5.4	92.2	4,739
Switzerland	95.8	4.2	92.0	11,234
Thailand	98.7	1.3	98.9	6,606
Tunisia	99.8	0.2	90.3	4,407
Turkey	98.5	1.5	98.2	4,848
United Arab Emirates	97.9	2.1	94.7	11,500
United Kingdom	94.6	5.4	86.1	12,659
United States	**94.6**	**5.4**	**88.9**	**6,111**
Uruguay	99.7	0.3	90.0	5,315
Vietnam	99.3	0.7	99.9	4,959
U.S. state education systems				
Connecticut	95.9	4.1	87.5	1,697
Florida	91.7	8.3	90.0	1,896
Massachusetts	95.6	4.4	90.0	1,723

NOTE: In calculating student participation rates, each student received a weight (student base weight) equal to the product of the school base weight—for the school in which the student was enrolled—and the reciprocal of the student selection probability within the school. Coverage of 15-year-old population refers to the extent to which the weighted participants covered the target population of all enrolled students in grades 7 and above. Coverage of national desired population refers to the extent to which the weighted participants covered the national population of 15-year-olds under the nonexcluded portion of the student sample. Overall student exclusion rate is the percentage of students excluded for intellectual or functional disabilities, or insufficient assessment language experience at either the school level or within schools. Weighted student participation after replacement refers to the sum of weights of students in original and replacement schools with PISA-assessed students and a student response rate of at least 50 percent over the sum of weights of students in responding original sample schools, responding replacement schools, and eligible refusing original sample schools. Italics indicate non-OECD countries and education systems. Results for Connecticut, Florida, and Massachusetts are for public school students only.
SOURCE: Organization for Economic Cooperation and Development (OECD), Program for International Student Assessment (PISA), 2012.

U.S. Nonresponse Bias Analysis

Of the 240 original sampled schools in the U.S. national sample, 207 were eligible (18 schools did not have any 15-year-olds enrolled, 6 had closed, and 9 were otherwise ineligible), and 139 agreed to participate. The weighted school response rate before replacement was 67 percent, requiring the United States to conduct a nonresponse bias analysis, which was used by the PISA consortium and the Organization for Economic Cooperation and Development (OECD) to evaluate the quality of the final sample.

A bias analysis was conducted in the United States to address potential problems in the data owing to school nonresponse. To compare PISA participating schools to the total eligible sample of schools, it was necessary to match the sample of schools to the sample frame to identify as many characteristics as possible that might provide information about the presence of nonresponse bias. Frame characteristics were taken from the 2008–09 Common Core of Data for public schools and from the 2009–10 Private School Universe Survey for private schools. The available school characteristics included affiliation (public or private), locale (city, suburb, town, rural), Census region, number of age-eligible students, total number of students, and percentage of various racial/ethnic groups (White, Black, Hispanic, non-Hispanic, Asian, American Indian or Alaska Native, Native Hawaiian/ Pacific Islander, and multiracial). The percentage of students eligible for free or reduced-price lunch was available for public schools only. The full text of the nonresponse bias analysis conducted for PISA 2012 will be included in the technical report released with the U.S. national dataset (Kastberg, Roey, Lemanski, Chan, and Murray forthcoming).

For original sample schools, participating schools had a higher mean percentage of Hispanic students than the total eligible sample of schools (21.1 versus 18.1 percent, respectively). Participating original sample schools also had a higher mean percentage of students eligible for free or reduced-price lunch than did the total eligible sample of schools (39.3 versus 36.1 percent, respectively). When all factors were considered simultaneously in a logistic regression analysis, only "town" (a territory inside an urban cluster with a core population between 25,000 and 50,000) was a significant predictor of participation. The percentage of students eligible for free or reduced-price lunch was not included in the logistic regression analysis as public and private schools were modeled together using only the variables available for all schools.[1]

For final sample schools (with substitutes), participating schools had a higher mean percentage of students eligible for free or reduced-price lunch than the total eligible sample of schools (38.4 versus 36.2 percent, respectively). When all factors were considered simultaneously in a logistic regression analysis (again with free or reduced-price lunch eligibility omitted), no variables were statistically significant predictors of participation.

With the inclusion of substitute schools and school nonresponse adjustments applied to the weights, only the percentage of students eligible for free or reduced-price lunch remained statistically significant. Specifically, the participating schools had a higher mean percentage of students eligible to receive free or reduced-price lunch than the total eligible sample of schools (38.4 versus 36.2 percent, respectively). However, there was not a statistically significant relationship between participating schools and the total frame of eligible schools for the percentage of students eligible for free or reduced-price lunch (38.4 versus 37.1 percent, respectively). We therefore conclude that, despite

[1] The nonresponse bias analysis was designed to measure the potential nonresponse bias for all participating schools, so no additional logistic regression was conducted using only public schools.

the tendency of schools with higher percentages of students eligible for free and reduced-price lunch to participate at a greater rate than other sampled schools, there is little evidence of resulting potential bias in the final sample. The multivariate regression analysis cannot be conducted after the school nonresponse adjustments are applied to the weights. The concept of nonresponse adjusted weights does not apply to the nonresponding units, and, thus, we cannot conduct an analysis that compares respondents with nonrespondents using nonresponse adjusted weights.

In sum, the investigation into nonresponse bias at the school level in the United States in PISA 2012 provides evidence that there is little potential for nonresponse bias in the PISA participating sample based on the characteristics studied. It also suggests that, while there is little evidence that the use of substitute schools reduced the potential for bias, it has not added to it. Moreover, the application of school nonresponse adjustments substantially reduced the potential for bias.

www.ed.gov

ies.ed.gov