Program for International Student Assessment (PISA)

Website: http://nces.ed.gov/Surveys/PISA/
Updated: August 2014

1. OVERVIEW

The Program for International Student Assessment (PISA) is a system of international assessments that measures 15-year-old students' capabilities in reading literacy, mathematics literacy, and science literacy every three years. PISA, first implemented in 2000, was developed and is administered under the auspices of the Organization for Economic Cooperation and Development (OECD), an intergovernmental organization of industrialized countries.[1] PISA 2012 was the fifth in this series of assessments; the next cycle of data is being collected in 2015. The PISA Consortium, a group of international organizations engaged by the OECD, is responsible for coordinating the study operations across countries and currently consists of the German Institute for Educational Research and the Educational Testing Service. The National Center for Education Statistics (NCES), in the Institute of Education Sciences at the U.S. Department of Education, is responsible for the implementation of PISA in the United States.

INTERNATIONAL ASSESSMENT OF 15-YEAR-OLDS:

Assesses literacy skills in the following areas:

- Reading literacy
- Mathematics literacy
- Science literacy

PISA was implemented in 43 countries and economies in the first cycle (32 in 2000 and 11 in 2002), 41 in the second cycle (2003), 57 in the third cycle (2006) and 75 in the fourth cycle (65 in 2009 and 10 in 2010). In PISA 2012, 65 countries and economies participated. The test is typically administered to between 4,500 and 10,000 students in each country/economy. Economies are regions of a country that participate in PISA separately from the whole country.

Purpose

PISA provides internationally comparative information on the reading, mathematics, and science literacy of students at an age that, for most economies, is near the end of compulsory schooling. The objective of PISA is to measure the "yield" of economies, or what skills and competencies students have acquired and can apply in reading, mathematics, and science to real-world contexts by age 15. The literacy concept emphasizes the mastery of processes, the understanding of concepts, and the application of knowledge and functioning in various situations. By focusing on literacy, PISA draws not only from school curricula but also from learning that may occur outside of school.

[1] Countries that participate in PISA are referred to as jurisdictions or economies throughout this chapter.

Components

Assessment. PISA is designed to assess 15-year-olds' performance in reading, mathematics, and science literacy. PISA 2012 also included a problem solving assessment, in which not all countries participated because of technical issues, and computer-based reading, mathematics, and financial literacy, which participating economies had the option of administering.

PISA 2015 will include collaborative problem solving and financial literacy assessments in addition to the core assessment subjects. In 2012, PISA was administered as a paper-and-pencil assessment; in 2015, PISA will be entirely computer-based. Each student takes a two-hour assessment. Assessment items include a combination of multiple-choice questions, closed- or short- response questions (for which answers are either correct or incorrect), and open-constructed response questions (for which answers can receive partial credit).

Questionnaires. Students complete a 30-minute questionnaire providing information about their backgrounds, attitudes, and experiences in school. In addition, the principal of each participating school completes a 30-minute questionnaire on school characteristics and policies. Teacher questionnaires will be a new addition to the PISA data collection starting in 2015.

Periodicity

PISA operates on a three-year cycle. Each PISA assessment cycle focuses on one subject in particular, although all three subjects are assessed every year. In 2000, PISA focused on reading literacy; in 2003, on mathematics literacy (including problem solving); and in 2006, on science literacy. In 2009, the focus was again on reading literacy, and PISA 2012 focused on mathematics (including problem solving and financial literacy). In 2015, PISA will focus on science literacy (including collaborative problem solving and financial literacy).

2. USES OF DATA

PISA provides valuable information for comparisons of student performance across jurisdictions and over time at the national level and for some jurisdictions at the subnational level. Performance in each subject area can be compared across jurisdictions in terms of:

- economies' mean scores;
- the proportion of students in each jurisdiction reaching PISA proficiency levels;
- the scores of economies' highest performing and lowest performing students;
- the standard deviation of scores in each jurisdiction; and
- other measures of the distribution of performance within jurisdictions.

PISA also supports cross-jurisdictional comparisons of the performance of some subgroups of students, including students grouped by sex, immigrant status, and socioeconomic status. PISA data are not useful for comparing the performance of racial/ethnic groups across jurisdictions because relevant racial/ethnic groups differ across jurisdictions. However, PISA datasets for the United States include information that can be used in comparing groups of students by race/ethnicity and school poverty level.

Contextual measures taken from student and principal questionnaires can be used to compare the educational contexts of 15-year-old students across jurisdictions. Caution should be taken, however, in attempting to interpret associations between measures of educational context and student performance. The PISA assessment is intended to tap factual knowledge and problem-solving skills that students learn over several years, whereas PISA contextual measures typically reference students' current school context. In the United States, for example, data collection occurs in the fall of the school year; therefore, contextual measures may apply to schools that children have attended for only 1 or 2 months.

Through the collection of comparable information across jurisdictions at the student and school levels, PISA adds significantly to the knowledge base that was previously available only from official national statistics.

3. KEY CONCEPTS

The types of literacy measured by PISA are defined as follows.

Reading literacy. An individual's capacity to understand, use, reflect on and engage with written texts, in order to achieve one's goals, to develop one's knowledge and potential, and to participate in society.

Mathematics literacy. An individual's capacity to identify and understand the role that mathematics plays in the world, make well-founded judgments, and use and engage with mathematics in ways that meet one's

needs as a constructive, concerned, and reflective citizen.

The PISA mathematics framework was updated for the 2012 assessment. The revised framework is intended to clarify the mathematics relevant to 15-year-old students, while ensuring that the items developed remain set in meaningful and authentic contexts, and defines the mathematical processes in which students engage as they solve problems. These processes, described above, are being used for the first time in 2012 as a primary reporting dimension. Although the framework has been updated, it is still possible to measure trends in mathematics literacy over time, as the underlying construct is intact.

Science literacy. An individual's scientific knowledge and the use of that knowledge to identify questions, acquire new knowledge, explain scientific phenomena, and draw evidence-based conclusions about science-related issues; an understanding of the characteristic features of science as a form of human knowledge and inquiry; an awareness of how science and technology shape our material, intellectual, and cultural environments; and a willingness to engage in science-related issues—and with the ideas of science—as a reflective citizen.

4. SURVEY DESIGN

The survey design for PISA data collections is discussed in this section.

Target Population

The desired PISA target population consisted of 15-year-old students attending public or private educational institutions located within the jurisdiction, in grades 7 through 12. Jurisdictions were to include 15-year-old students enrolled either full time or part time in an educational institution, in a vocational training or related type of educational program, or in a foreign school within the jurisdiction (as well as students from other jurisdictions attending any of the programs in the first three categories). It was recognized that no testing of persons schooled in the home, workplace, or out of the jurisdiction occurred; therefore, these students were not included in the international target population.

The operational definition of an age population depends directly on the testing dates. International 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 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, for PISA 2012. In the United States, students born between July 1, 1996, and June 30, 1997, were eligible to participate in PISA 2012.

The U.S. PISA 2012 national school sample consisted of 240 schools. 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.

International Sample Design

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 economy.

A minimum of 4,500 students from a minimum of 150 schools was required in each country. 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 42 students was to be selected in an equal probability sample unless fewer than 42 students age 15 were available (in which case all 15-year-old students were selected).

International within-school exclusion rules for students were specified as follows:

- *Students with functional disabilities.* These were students with a moderate to severe permanent physical disability such that they could not perform in the PISA testing environment.

- *Students with intellectual disabilities.* These were students with a mental or emotional

disability who had been tested as cognitively delayed or who were considered in the professional opinion of qualified staff to be cognitively delayed such that they could not perform in the PISA testing situation.

- *Students with insufficient language experience.* These were students who met the three criteria of (1) not being a native speaker in the assessment language, (2) having limited proficiency in the assessment language, and (3) having received less than a year of instruction in the assessment language. In the United States, English was the exclusive language of the assessment.

A school attended only by students who would be excluded for functional, intellectual, or linguistic reasons was considered a school-level exclusion. International exclusion rules for schools allowed for schools in remote regions or very small schools to be excluded. School-level exclusions for inaccessibility, feasibility, or other reasons were required to cover fewer than 0.5 percent of the total number of students in the international PISA target population. International guidelines state that no more than 5 percent of a jurisdiction's desired national target population should be excluded from the sample.

A minimum of 150 schools (or all schools, if there were fewer than 150 in a participating jurisdiction) had to be selected in each jurisdiction. Within each participating school, a sample of the PISA-eligible students was selected with equal probability. In total, a minimum sample size of 4,500 assessed students was to be achieved in each jurisdiction. If a jurisdiction had fewer than 4,500 eligible students, then the sample size was the *national defined target population*. The national defined target population included all eligible students in the schools that were listed in the school sampling frame.

Response Rate Targets

School response rates. The PISA international guidelines for the 2012 assessment required that jurisdictions achieve an 85 percent school response rate. However, while stating that each jurisdiction must make every effort to obtain cooperation from the sampled schools, the requirements also recognized that this is not always possible. Thus, it was allowable to use substitute, or replacement, schools as a means to avoid loss of sample size associated with school nonresponse. The international guidelines stated that at least 65 percent of participating schools must be from the original sample. Economies were only allowed to use replacement schools (selected during the sampling process) to increase the response rate once the 65 percent benchmark had been reached.

Each sampled school was to be assigned two replacement schools in the sampling frame. If the original sampled school refused to participate, a replacement school was asked to participate. One sampled school could not substitute for another sampled school, and a given school could only be assigned to substitute for one sampled school. A requirement of these substitute schools was that they be in the same explicit stratum as the original sampled school. The international guidelines define the response rate as the number of participating schools (both original and replacement schools) divided by the total number of eligible original sampled schools.[2]

Student response rates. A minimum response rate of 80 percent of selected students across participating schools was required. 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.

Within each school, a student response rate of 50 percent was required for a school to be regarded as participating: the overall student response rate was computed using only students from schools with at least a 50 percent response rate. Weighted student response rates were used to determine if this standard was met; each student's weight was the reciprocal of his or her probability for selection into the sample.

Sample Design in the United States

The design of the U.S. school sample for PISA 2012 was developed to achieve each of the international requirements set forth in the PISA sampling manual. The U.S. school sample was a stratified systematic sample, consisting of two stages, and was intended to approximate a self-weighting sample of students, with each 15-year-old student having equal probability of being selected. In the first stage, schools were selected with a probability proportionate to the school's estimated enrollment of 15-year-olds. In the second

[2] The calculation of response rates described here is based on the formula stated in the international guidelines and is not consistent with NCES standards. A more conservative way to calculate response rates would be to include participating replacement schools in the denominator as well as in the numerator and to add replacement schools that were hard refusals to the denominator.

stage, a sample of 50 students was selected from each school in an equal probability sample, regardless of size (all eligible students were selected if there were fewer than 50). The United States set a target cluster size (TCS) of 50 students per school in order to achieve the required student yield of 35 assessed students per school (taking into account student exclusions and absences). The TCS for the main study was slightly larger than the TCS used on PISA 2009 in the United States to account for the financial literacy assessment. Out of the 50 students, 42 were sampled to take the paper-based mathematics, science, and reading literacy assessment. Out of these 42 students, 20 were subsampled to also take the computer-based assessment. The remaining eight students were sampled to take the financial literacy assessment. If fewer than 50 age-eligible students were enrolled in a school, all 15-year-old students in that school were selected. The U.S. national TCS and student sampling plans were approved by the international consortium. Within each stratum, the frame was implicitly stratified (i.e., 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); first three digits of the zip code; combined percentage of Black, Hispanic, Asian, Pacific Islander, and American Indian/Alaska Native students (above or below 15 percent); and estimated enrollment of 15-year-olds.

The 2012 PISA sampling employed techniques to undersample very small schools (those with fewer than twenty-one 15-year-olds) and to minimize overlap with the High School Longitudinal Study of 2012 (HSLS:12), a U.S. education study with data collection conducted in fall 2012. If any PISA substitute school overlapped with an originally sampled or first substitute HSLS school, the substitute was not to be contacted for PISA. Under this rule, none of the schools were eliminated from the list of PISA substitute schools.

In 2012, as in each cycle since 2003, schools were selected in the first stage with probability proportional to size (PPS) sampling, and students were sampled in the second stage, yielding overall equal probabilities of selection. Comparatively in PISA 2000, the U.S. school sample had a three-stage design, the first of which was the selection of a sample of geographic primary sampling units (PSUs). The change to a two-stage model was made in PISA 2003 to reduce the design effects observed in the 2000 data and to minimize respondent burden on individual districts by distributing the response burden of the study across districts as much as possible.

Once the list of students was received from a school, it was formatted for importing into KeyQuest, the sampling and data management software provided by ACER. KeyQuest was used to manage the sample, draw the student sample, track participation, and produce verification reports used to clean the data in preparation for submitting the data file to ACER.

A list of schools for the U.S. sample was prepared using data from the 2008-09 Common Core of Data (CCD) and the 2009-10 Private School Universe Survey (PSS), two NCES surveys with the most current data at the time of the PISA frame construction. The U.S. school sample for PISA 2012 consisted of 240 schools containing at least one 7th through 12th grade class—194 large schools with at least 50 estimated eligible students, 11 moderately small schools with between 25 and 50 estimated eligible students and 16 very small schools with less than 25 but greater than 2 estimated eligible students, and 19 very small schools with estimated eligible enrollment of less than equal to 2 eligible students. Eligible schools in the PISA 2012 school frame included 207 of the 240 original sampled schools in the U.S. national sample (18 schools did not have any 15-year-olds enrolled, 6 had closed, and 9 were otherwise ineligible), and 139 agreed to participate.

Assessment Design

Test scope and format. In PISA 2012, the three subject domains were tested, with mathematics as the major domain and reading and science as the minor domains. Every student answered mathematics items, and some students answered reading items, science items, or both reading and science items.

The development of the PISA 2012 assessment instruments was an interactive process among the PISA Consortium, various expert committees, and OECD members. The assessment included items submitted by participating jurisdictions and items developed by the consortium's test developers. Representatives of each jurisdiction reviewed the items for possible bias and for relevance to PISA's goals. The intention was to reflect in the assessment the national, cultural, and linguistic variety of the OECD jurisdictions. Following a field trial that was conducted in most jurisdictions, test developers and expert groups considered a variety of aspects in selecting the items for the main study: (a) the results from the field trial, (b) the outcome of the item review from jurisdictions, and (c) queries received about the items.

PISA 2012 was a paper-and-pencil assessment. Approximately half of the items were multiple choice, about 20 percent were closed- or short-response items (for which students wrote an answer that was simply either correct or incorrect), and about 30 percent were open constructed-response items (which were graded

by trained scorers using an international scoring guide and could be assigned partial credit).

Multiple-choice items were either (a) standard multiple choice, with a limited number (usually four) of responses from which students were required to select the best answer; or (b) complex multiple choice, which presented several statements, each of which required students to choose one of several possible responses (true/false, correct/incorrect, etc.). Closed- or short-response items included items that required students to construct their own responses from a limited range of acceptable answers or to provide a brief answer from a wider range of possible answers, such as mathematics items requiring a numeric answer, and items requiring a word or short phrase. Open constructed-response items required more extensive writing, or showing a calculation, and frequently included some explanation or justification. Pencils, erasers, rulers, and (in some cases) calculators were provided.

In 2012, computer-based assessments in mathematics and reading were offered as optional assessments for participating economies. Thirty-two economies, including the United States, chose to administer them. In these economies, a subset of students who took the paper-based assessment also took an additional computer-based assessment. Although the paper-based assessment items and the computer-based assessment items were derived from the same frameworks, there was no overlap in the assessment items between the two assessment modes. The interactive nature of computer-based assessment allowed PISA to assess students in novel contexts that are not possible with a traditional paper-based format.

Test design. The PISA 2012 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 economies 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 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, problem solving and/or financial literacy items.

For 2012, a subset of students who took the paper-based assessment also took a 40-minute computer-based assessment. The computer-based assessment consisted of 168 problem-solving items, 164 mathematics items, and 144 reading items allocated to 24 assessment 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. 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.

Data Collection and Processing

PISA 2012 was coordinated by the OECD and managed at the international level by the PISA Consortium. PISA is implemented in each education system by a National Project Manager (NPM). In the United States, the NPM works with a national data collection contractor to implement procedures prepared by the PISA Consortium and agreed to by the participating jurisdictions. In 2012, the U.S. national data collection contractor was Westat as well as a subcontractor, Pearson. A steering committee also gave input on the dissemination and development of PISA in the United States.

The 2012 PISA multicycle study was again collaboration between the governments of participating countries, the Organization for Economic Cooperation and Development (OECD), and a consortium of various international organizations, referred to as the PISA Consortium. This consortium in 2012 was led by the Australian Council for Educational Research (ACER) and includes the German Institute for International Educational Research (DIPF), the German Social Sciences Infrastructure Services' Centre for Survey Research and Methodology (GESIS-ZUMA), the University of Maastricht's Research Centre for Education and the Labour Market (ROA), the U.S. research company Westat, the International Association for the Evaluation of Educational Achievement (IEA),

and the Belgian firm CapStan. There have been changes to the PISA consortium in charge of administering the 2015 data collection.

Reference dates. Each economy 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 economies conducted testing from March through August 2012. 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.

Incentive. School packages were mailed to principals in mid-September with phone contact from recruiters beginning a few days after the mailing. As part of the PISA 2012 school recruitment strategy, the materials included a description of school and student incentives. Schools and school coordinators were each paid $200, and students received $25 and 4 hours of community service for participating in the paper-based session and an additional $15 if they were selected and participated in the computer-based assessment.

Data collection. Each economy collected its own data. The PISA consortium emphasizes the implementation of standardized procedures in all jurisdictions (data collection followed a manual developed by the PISA Consortium). Professional staff, trained in the international guidelines, were responsible for test administration. School staff members were only responsible for specifying parental consent requirements, listing students, and providing testing space.

To ensure quality, test administrators were observed in a sample of schools in each jurisdiction by a PISA Quality Monitor (PQM). PQMs were engaged by the PISA Consortium itself. Observed schools were chosen jointly by the PISA Consortium and the PQMs; in the United States, a total of 7 schools were observed by the PQM. The main responsibility of the PQM was to record the extent to which testing procedures in schools were implemented in accordance with the standard test administration procedures. In U.S. schools, the PQM's observations indicated that international procedures for data collection were applied consistently.

The students were randomly assigned one of 17 test booklets, which test administrators distributed. U.S. students who took the mathematics assessment in PISA 2012 were allowed to use, and were provided, calculators.

Scoring. A substantial portion of the PISA 2012 assessment was devoted to open constructed-response items. The process of scoring these items is an important step in ensuring the quality and comparability of the PISA data. Detailed guidelines were developed for the scoring guides themselves, training materials to recruit scorers, and workshop materials used for the training of national scorers. Prior to the national training, the PISA Consortium organized international training sessions to present the material and train scoring coordinators from the participating jurisdictions, who in turn trained the national scorers.

For each test item, the scoring guides described the intent of the question and how to code students' responses. This description included the credit labels—full credit, partial credit, or no credit—attached to the possible categories of response. Also included was a system of double-digit coding for some mathematics and science items, where the first digit represented the score and the second digit represented the different strategies or approaches that students used to solve the problem. The second digit generated national profiles of student strategies and misconceptions. In addition, the scoring guides included real examples of students' responses accompanied by a rationale for their classification for purposes of clarity and illustration.

To examine the consistency of this marking process in more detail within each jurisdiction (and to estimate the magnitude of the variance components associated with the use of scorers), the PISA Consortium generated an inter-rater reliability report on a subsample of assessment booklets. The results of the homogeneity analysis showed that the marking process of items is largely satisfactory and that on average countries are more or less reliable in the coding of the open-ended responses.

For the PISA 2012, 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).

Data entry and verification. In PISA 20012 each jurisdiction was responsible for entering data into data files following a common international format. Variables could be added or deleted as needed for different national options; approved adaptations to

response categories could also be accommodated. Student response data were entered directly from the test booklets and questionnaires using specialized software that allowed the data files to be merged into KeyQuest and facilitated the checking and correction of data through various data consistency checks. After these checks, the data were sent to ACER for data cleaning; there, the data were checked to ensure they followed the international structure, the identification system was reviewed, single case problems were corrected manually, and standard data cleaning procedures were applied to the questionnaire files.

During data cleaning, analysts identified as many anomalies and inconsistencies as possible, and through a process of extensive discussion between each national center and ACER, an effort was made to correct and resolve all data issues. After this, ACER compiled background univariate statistics and performed preliminary classical and Rasch item analysis.

Estimation Methods

Weighting. The use of sampling weights is necessary for the computation of statistically sound, nationally representative estimates. Adjusted survey weights account 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.

The internationally defined weighting specifications for PISA 2012 included base weights and adjustments for nonresponse. The school base weight was defined as the reciprocal of the school's probability of selection. (For substitute 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 student selected from within a school.

These base weights were then adjusted for school and student nonresponse. The school nonresponse adjustment was done individually for each jurisdiction using implicit and explicit strata defined as part of the sample design. In the case of the United States, two variables were used: school control and Census region. The student nonresponse adjustment was done within cells based first on students' final school nonresponse rate and their explicit stratum; within that, grade and gender were used.

All PISA 2012 analyses were conducted using these adjusted sampling weights.

Scaling. There were 13 test booklets, each containing a slightly different subset of items, in the PISA 2012 design. Economies that participated in the financial literacy assessment had a total of 17 booklets each. Each student completed one test booklet. The fact that each student completed only a subset of items means that classical test scores, such as the percent correct, are not accurate measures of student performance. Instead, scaling techniques were used to establish a common scale for all students. In PISA 2009, item response theory (IRT) was used to estimate average scores in each jurisdiction for science, mathematics, and reading literacy, as well as for three reading literacy subscales: integrating and interpreting, accessing and retrieving, and reflecting and evaluating. Subscale scores were not available for mathematics literacy or science literacy for 2009 because not all students answered science and/or mathematics items.

IRT identifies patterns of response and uses statistical models to predict the probability of a student answering an item correctly as a function of his or her proficiency in answering other questions. PISA 2009 used a mixed coefficients multinomial logit IRT model. This model is similar in principle to the more familiar two-parameter logistic IRT model. With the multinomial logit IRT model, 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.

IRT was used for PISA 2012 also, to estimate average scores for mathematics, science, and reading literacy for each economy, as well as for three mathematics process and four mathematics content scales. For economies participating in the financial literacy assessment and the computer-based assessment, these assessments will be scaled separately and assigned separate scores.

Plausible values. Scores for students are estimated as plausible values because each student completed only a subset of items. These values represent the distribution of potential scores for all students in the population with similar characteristics and identical patterns of item response. It is important to recognize that plausible values are not test scores and should not be treated as such. Plausible values are randomly drawn from the distribution of scores that could be reasonably assigned to each individual. As such, the plausible values contain random error variance components and are not optimal as scores for individuals. Five plausible values were estimated for each student for each scale in PISA 2012. Thus, statistics describing performance on the PISA science, reading, and mathematics literacy scales are based on plausible values.

If an analysis is to be undertaken with one of these cognitive scales, then (ideally) the analysis should be

undertaken five times, once with each of the five relevant plausible value variables. The results of these five analyses are averaged; then, significance tests that adjust for variation between the five sets of results are computed.

Imputation. Missing background data from student and principal questionnaires are not imputed for PISA 2009 reports. PISA 2012 also did not impute missing information for questionnaire variables.

In general, item response rates for variables discussed in NCES PISA reports exceed the NCES standard of 85 percent.

Measuring trends. Although the PISA 2012 framework was updated, it is still possible to measure trends in mathematics literacy over time, as the underlying construct is intact. For specific trends in performance results, please see the NCES PISA website (http://nces.ed.gov/surveys/pisa/pisa2012/). Reading literacy scales used in PISA 2000, PISA 2003, PISA 2006, PISA 2009, and PISA 2012 are directly comparable, which means that the value of 500 in PISA 2012 has the same relative meaning as it did in PISA 2000, PISA 2003, PISA 2006, and PISA 2009. However, for PISA 2003, the mathematics assessment underwent major development work and was broadened to include four sub-domains; only two of these appeared in PISA 2000. As such, mathematics literacy scales are only comparable between PISA 2003, PISA 2006, PISA 2009, and PISA 2012. Likewise, PISA 2006 was the first major assessment of science literacy; thus, the science literacy scale in PISA 2006 is only directly comparable with PISA 2009 and PISA 2012.

The PISA 2000, 2003, 2006, 2009, and 2012 assessments of reading, mathematics, and science are linked assessments. That is, the sets of items used to assess each domain in each year include a subset of common items; these common items are referred to as link items. In PISA 2000 and PISA 2003, there were 28 reading items, 20 math items, and 25 science items that were used in both assessments. The same 28 reading items were retained in 2006 to link the PISA 2006 data to PISA 2003, The PISA 2009 assessment included 26 of these 28 reading items and a further 11 reading items from PISA 2000, not used since that administration, were also included in PISA 2009. The PISA 2012 assessment included 37 of these link items from 2009 as well as an additional 7 items included in 2009 to establish the reading trend scale. In mathematics, 48 math items from PISA 2003 were used in PISA 2006; PISA 2009 included 35 of the 48 mathematics items that were used in PISA 2006, and of these, 34 were used in PISA 2012. For the science assessment, 14 items were common to PISA 2000 and PISA 2006, and 22 items were common to PISA 2003 and PISA 2006. The science assessment for PISA 2012 consisted of 53 items that were used in PISA 2009 and 2006.

To establish common reporting metrics for PISA, the difficulty of the link items, measured on different occasions, is compared. Using procedures that are detailed in the *PISA 2012 Technical Report*, the comparison of item difficulty on different occasions is used to determine a score transformation that allows the reporting of the data for a particular subject on a common scale. The change in the difficulty of the individual link items is used in determining the transformation; as a consequence, the sample of link items that has been chosen will influence the choice of transformation. This means that if an alternative set of link items had been chosen, the resulting transformation would be slightly different. The consequence is an uncertainty in the transformation due to the sampling of the link items, just as there is an uncertainty in values such as jurisdiction means due to the use of a sample of students.

Future Plans

The next cycle of PISA data collection will take place in 2015.

(http://nces.ed.gov/Surveys/PISA/)

5. DATA QUALITY AND COMPARABILITY

A comprehensive program of continuous quality monitoring was central to ensuring full, valid implementation of the PISA procedures and the recording of deviations from these procedures. Quality monitors from the PISA Consortium visited a sample of schools in every jurisdiction to ensure that testing procedures were carried out in a consistent manner. The purpose of quality monitoring is to observe and record the implementation of the described procedures; therefore, the field operations manuals provided the foundation for all the quality monitoring procedures.

The manuals that formed the basis for the quality monitoring procedures were the PISA Consortium data collection manual and the PISA data management manual. In addition, the PISA data were verified at several points starting at the time of data entry.

Despite the efforts taken to minimize error, as with any study, PISA has limitations that researchers should take into consideration. This section contains a discussion of two possible sources of error in PISA: sampling and nonsampling errors.

Sampling Error

Sampling errors occur when a discrepancy between a population characteristic and the sample estimate arises because not all members of the target 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 2011-12 school year was just one of many possible samples that could have been selected. Therefore, estimates produced from the PISA 2012 sample may differ from estimates that would have been produced had another sample of students been selected. 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 owing to sampling when estimating a statistic. The approach used for calculating sampling variances in PISA is Fay's method of balanced repeated replication (BRR). 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 (SRS) assumptions for non-SRS data. Thus, the standard errors reported in PISA can be used as a measure of the precision expected from this particular sample.

Nonsampling Error

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, the sampling frame in the United States was limited to regular public and private schools in the 50 states and the District of Columbia and cannot be used to represent Puerto Rico or other jurisdictions (e.g., other U.S. territories and DoD schools overseas). 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, response differences related to the particular time the survey was conducted, and mistakes in data preparation.

In general, it is difficult to identify and estimate either the amount of nonsampling error or how much bias it causes. In PISA 2012, efforts were made to prevent such errors from occurring and to compensate for them when possible. For example, the design phase entailed a field test that evaluated items as well as the implementation procedures for the survey. One type of nonsampling error that may be present in PISA is respondent bias, which occurs when respondents systematically misreport (intentionally or not) information in a study; a potential source of respondent bias in this survey was social desirability bias. For example, students may overstate their parents' educational attainment or occupational status. If there were no systematic differences among specific groups under study in their tendency to give socially desirable responses, then comparisons of the different groups would accurately reflect differences among groups. Readers should be aware that respondent bias may be present in this survey as in any survey; however, it is not possible to state precisely how such bias may affect the results.

Coverage error. Every National Project Manager (NPM) was required to define and describe their jurisdiction's national desired target population and explain how and why it might deviate from the international target population. Any hardships in accomplishing complete coverage were specified, discussed, and approved (or not) in advance. Where the national desired target population deviated from full national coverage of all eligible students, the deviations were described and enrollment data provided to measure how much that coverage was reduced. School-level and within-school exclusions from the national desired target population resulted in a national defined target population corresponding to the population of students recorded in each jurisdiction's school sampling frame.

In PISA 2009, the United States reported 82 percent coverage of the 15-year-old population and 95 percent coverage of the national desired target population (OECD 2010). The United States reported a 5.2 percent overall exclusion rate, which was higher than the internationally acceptable exclusion rate of 5 percent. However, when language exclusions were accounted for (i.e., removed from the overall exclusion rate), the United States no longer had an exclusion rate greater than 5 percent. For PISA 2012, 95 percent coverage of the national desired target population was achieved.

Nonresponse error. Nonresponse error results from nonparticipation of schools and students. School nonresponse, without replacement schools, will lead to the underrepresentation of students from the type of school that did not participate, unless weighting adjustments are made. It is also possible that only a part of the eligible population in a school (such as those 15-year-olds in a single grade) was represented by the school's student sample; this also requires weighting to compensate for the missing data from the omitted grades. Student nonresponse within participating schools occurred to varying extents. Students who could not be given achievement test scores but were not excluded for linguistic or disability reasons, will be

underrepresented in the data unless weighting adjustments are made.

Unit nonresponse. Of the 240 original sampled schools in the PISA 2012 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. However, 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.

Table PISA-1. U.S. weighted school and student response rates: PISA 2012

	Weighted response rate (percent)
School	
Before replacement	67
After replacement	77
Student	89

SOURCE: Kelly, D., Xie, H., Nord, C.W., Jenkins, F., Chan, J.Y., and Kastberg, D. (2013). *Performance of U.S. 15-Year-Old Students in Mathematics, Science, and Reading Literacy in an International Context: First Look at PISA 2012* (NCES 2014-024). U.S. Department of Education. Washington, DC: National Center for Education Statistics. (pp. B-3, B-5)

A total of 6,110 students in the United States were sampled for the PISA 2012 assessment. The overall student exclusion rate for the United States was 5.3 percent, with an overall weighted student participation rate after replacement of 89 percent.

For PISA 2012, a bias analysis was conducted in the United States to address potential problems in the data owing to school nonresponse; however, the investigation into nonresponse bias at the school level in the United States in PISA 2012 provided evidence that there is little potential for nonresponse bias in the PISA participating sample based on the characteristics studied. 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.

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). All factors were then considered simultaneously in a logistic regression analysis, and 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. 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); this means that despite 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.

Measurement error. Measurement error is introduced into a survey when its test instruments do not accurately measure the knowledge or aptitude they are intended to assess.

Data Comparability

A number of international comparative studies already exist to measure achievement in mathematics, science, and reading, including the Trends in International Mathematics and Science Study (TIMSS) and the Progress in International Reading Literacy Study (PIRLS). The Adult Literacy and Lifeskills Survey (ALL) was last conducted in 2003 and measured the literacy and numeracy skills of adults. A new study, the Program for the International Assessment of Adult Competencies (PIAAC), was administered for the first time in 2011 and assessed the level and distribution of adult skills required for successful participation in the economy of participating jurisdictions. In addition, the United States has been conducting its own national surveys of student achievement for more than 35 years through the National Assessment of Educational Progress (NAEP). PISA differs from these studies in several ways.

Content. PISA is designed to measure literacy broadly, whereas studies such as TIMSS and NAEP have a stronger link to curricular frameworks and seek to measure students' mastery of specific knowledge, skills, and concepts. The content of PISA is drawn from broad content areas (e.g., space and shape in mathematics) in contrast to more specific curriculum-based content, such as geometry or algebra. For example, with regard to the reading assessment, PISA must contain passages applicable to a wide range of cultures and languages, making it unlikely that the passages will be intact, existing texts.

Tasks. PISA also differs from other assessments in that it emphasizes the application of reading, mathematics, and science literacy to everyday situations by asking students to perform tasks that involve interpretation of real-world materials as much as possible. A study comparing the PISA, NAEP, and TIMSS mathematics assessments found that the mathematics topics addressed by each assessment are similar, although PISA places greater emphasis on data analysis and less on algebra than does either NAEP or TIMSS. However, it is in how that content is presented that makes PISA different. PISA uses multiple-choice items less frequently than NAEP or TIMSS, and it contains a higher proportion of items reflecting moderate to high mathematical complexity than do those two assessments.

An earlier comparative analysis of the PISA, TIMSS, and NAEP mathematics and science assessments also found differences between PISA and the other two studies. In science, it found that more items in PISA built connections to practical situations and required students to demonstrate multistep reasoning and fewer items used a multiple-choice format than in NAEP or TIMSS. In mathematics, it found that more items in PISA than in NAEP or TIMSS were set in real-life situations or scenarios, required multistep reasoning, and required interpretation of figures and other graphical data. These tasks reflect the underlying assumption of PISA: as 15-year-olds begin to make the transition to adult life, they need to know how to read or use particular mathematical formulas or scientific concepts, as well as how to apply this knowledge and these skills in the many different situations they will encounter in their lives.

Age-based sample. In contrast with TIMSS and PIRLS, which are grade-based assessments, PISA's sample is based on age. TIMSS assesses fourth- and eighth-graders, while PIRLS assesses only fourth-graders. The PISA sample, however, is drawn from 15-year-old students, regardless of grade level. The goal of PISA is to represent outcomes of learning rather than outcomes of schooling. By placing the emphasis on age, PISA intends to show not only what 15-year-olds have learned in school in a particular grade, but outside of school as well as over the years. PISA thus seeks to show the overall yield of an economy and the cumulative effects of all learning experience. Focusing on age 15 provides an opportunity to measure broad learning outcomes while all students are still required to be in school across the many participating jurisdictions. Finally, because years of education vary among jurisdictions, choosing an age-based sample makes comparisons across jurisdictions somewhat easier.

6. CONTACT INFORMATION

For content information about PISA, contact

Patrick Gonzales
Phone: 415-920-9229
Email: patrick.gonzales@ed.gov

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

7. METHODOLOGY AND EVALUATION REPORTS

Most of the technical documentation for PISA is published by the OECD. The U.S. Department of

Education, NCES, is the source of several additional references listed below.

General

Baldi, S., Jin, Y., Skemer, M., Green, P., Herget, D., and Xie, H. (2007). *Highlights From PISA 2006: Performance of U.S. 15-Year-Olds in Science and Mathematics Literacy in an International Context* (NCES 2008-016). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC. Available at http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2008016.

Fleischman, H.L., Hopstock, P.J., Pelczar, M.P., and Shelley, B.E. (2010). *Highlights From PISA 2009: Performance of U.S. 15-Year-Old Students in Reading, Mathematics, and Science Literacy in an International Context* (NCES 2011-004). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC. Available at http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2011004.

Kastberg D., Roey S., Lemanski N. Chan J. Y., and Murray G. (2014). *Technical Report and User Guide for the Program for International Student Assessment (PISA): 2012 Data Files and Database with U.S.-Specific Variables* (NCES 2014-025). U.S. Department of Education. Washington, DC: National Center for Education Statistics. Available at http://nces.ed.gov/pubs2014/2014025.pdf.

Hopstock, P. and Pelczar, M. (2011). *Technical Report and User's Guide for the Program for International Student Assessment (PISA): 2009 Data Files and Database with U.S. Specific Variables* (NCES 2011-025). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC: U.S. Government Printing Office. Available at http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2011025.

Kelly, D., Xie, H., Nord, C.W., Jenkins, F., Chan, J.Y., and Kastberg, D. (2013). *Performance of U.S. 15-Year-Old Students in Mathematics, Science, and Reading Literacy in an International Context: First Look at PISA 2012* (NCES 2014-024). U.S. Department of Education. Washington, DC: National Center for Education Statistics. Available at http://nces.ed.gov/pubs2014/2014024rev.pdf.

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). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC. Available at http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2005003.

Organization for Economic Cooperation and Development (OECD). (2005). *PISA 2003 Data Analysis Manual*. Paris: Author.

Organization for Economic Cooperation and Development (OECD). (2007). *PISA 2006*: *Science Competencies for Tomorrow's World. Volume 1: Analysis.* Paris: Author.

Organization for Economic Cooperation and Development (OECD). (2009). *PISA 2009 Assessment Framework - Key Competencies in Reading, Mathematics and Science*. Paris: Author.

Organization for Economic Cooperation and Development (OECD). (2010). *PISA 2009 Results: What Students Know and Can Do - Performance in Reading, Mathematics and Science (Volume I).* Paris: Author.

Organization for Economic Cooperation and Development (OECD). (2012). *PISA 2009 Technical Report.* Paris: Author.

Survey Design

Green, P., Herget, D., and Rosen, J. (2009). *User's Guide for the Program for International Student Assessment (PISA): 2006 Data Files and Database With United States Specific Variables* (NCES 2009-055). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC. Available at http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2009055.

Organization for Economic Cooperation and Development (OECD). (2009). *PISA 2006 Technical Report*. Paris: Author.

Organization for Economic Cooperation and Development (OECD). (2010). *PISA 2009 Results: What Students Know and Can Do—Student Performance in Reading, Mathematics and Science (Volume I).* Paris: Author.

Organization for Economic Cooperation and Development (OECD). (2013). *PISA 2012 Assessment Framework—Key Competencies in Reading, Mathematics and Science*. Paris: Author.

Organization for Economic Cooperation and Development (OECD). (2013), *PISA 2012 Assessment and Analytical Framework: Mathematics, Reading, Science, Problem Solving and Financial Literacy*, OECD Publishing. Available at http://www.oecd.org/pisa/pisaproducts/PISA%202012%20framework%20e-book_final.pdf.

Organization for Economic Cooperation and Development (OECD). (2014), *PISA 2012 Technical Report*. Available at http://www.oecd.org/pisa/pisaproducts/pisa2012technicalreport.htm

PISA Project Consortium. (2005a). *Main Study National Project Manager's Manual.* Retrieved November 14, 2008, from https://mypisa.acer.edu.au/images/mypisadoc/opmanual/pisa2003_national_project_manager_manual.pdf.

PISA Project Consortium. (2005b). *School Sampling Preparation Manual: PISA 2006 Main Study.* Retrieved November 14, 2008, from https://mypisa.acer.edu.au/images/mypisadoc/opmanual/pisa2003_school_sampling_manual.pdf.

PISA Project Consortium. (2005c). *PISA 2006 Main Study Test Administrator's Manual.* Retrieved November 14, 2008, from https://mypisa.acer.edu.au/images/mypisadoc/opmanual/pisa2006_test_administrator_manual.pdf.

PISA Project Consortium. (2005d). *PISA 2006 Main Study School Coordinator's Manual.* Retrieved November 14, 2008, from https://mypisa.acer.edu.au/images/mypisadoc/opmanual/pisa2006_school_coordinators_manual.pdf.

Data Quality and Comparability

Dossey, J.A., McCrone, S.S., and O'Sullivan, C. (2006). *Problem Solving in the PISA and TIMSS 2003 Assessments* (NCES 2007-049). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC. Available at http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2007049.

Neidorf, T.S., Binkley, M., Gattis, K., and Nohara, D. (2006). *Comparing Mathematics Content in the National Assessment of Educational Progress (NAEP), Trends in International Mathematics and Science Study (TIMSS), and Program for International Student Assessment (PISA) 2003 Assessments* (NCES 2006-029). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC. Available at http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2006029.

Nohara, D. (2001). *A Comparison of the National Assessment of Educational Progress (NAEP), the Third International Mathematics and Science Study Repeat (TIMSS-R), and the Programme for International Student Assessment (PISA)* (NCES 2001-07). National Center for Education Statistics, U.S. Department of Education. Washington, DC. Available at http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=200107.

Stephens, M., and Coleman, M. (2007). *Comparing PIRLS and PISA With NAEP in Reading, Mathematics, and Science* (Working Paper). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC. Retrieved November 14, 2008, from http://nces.ed.gov/Surveys/PISA/pdf/comppaper12082004.pdf.