Program for International Student Assessment (PISA)

Website: https://nces.ed.gov/Surveys/PISA/
Updated: August 2020

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 2018 was the seventh in this series of assessments; the next cycle of data will be collected in 2021. 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.

PISA was implemented in 43 countries and education systems in the first cycle (32 in 2000 and 11 in 2002), 41 in the second cycle (2003), 57 in the third cycle (2006), 75 in the fourth cycle (65 in 2009 and 10 in 2010), 65 in the fifth cycle (2012), and 73 in the sixth cycle (2015). In PISA 2018, 79 countries and education systems participated. The test is typically administered to between 4,500 and 10,000 students in each country/education system. Education systems 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 education systems, is near the end of compulsory schooling. The objective of PISA is to measure the "yield" of education systems, 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.

Components
Assessment. PISA is designed to assess 15-year-olds' performance in reading, mathematics, and science literacy. PISA 2003 and PISA 2012 included an optional problem-solving assessment; in PISA 2012, not all countries participated in this assessment. PISA 2012 also introduced computer-based assessments for both reading and mathematics, as well as a paper-based financial literacy, assessment which participating education systems had the option of administering.

INTERNATIONAL ASSESSMENT OF 15-YEAR-OLDS:

Assesses literacy skills in the following areas:

- Reading literacy
- Mathematics literacy
- Science literacy

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

In 2015, PISA included an optional financial literacy assessment as well as a collaborative problem solving assessment that was different from the problem solving domain administered in 2003 and 2012. The assessments for all subjects were delivered via computer, and paper-based assessments were provided for countries that were not able not able to test the students by computer. PISA 2018 included an optional financial literacy assessment and also introduced a global competence assessment. Similar to 2015, all subjects were assessed through a computer-based assessment in a majority of countries and a paper-based assessment was provided for countries that were unable to administer a computer-based assessment. Each student took 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). In addition, PISA 2018 introduced multistage adaptive testing in reading assessment. Instead of using fixed, predetermined test booklets, the reading assessment given to each student was dynamically determined based on how student performed in prior stages (Core Stage, Stage 1 and Stage 2).

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 45 minute questionnaire on the school's demographics and learning environment. Teacher questionnaires were added in PISA 2015; up to 25 teachers per school completed 30-minute questionnaires on teaching practices, beliefs about teaching, and their qualifications and backgrounds.

In order to keep PISA as inclusive as possible and to keep the exclusion rate down, the United States the UH ('*Une Heure*') instrument designed for students with special education needs. The UH instrument was available to special education needs students within mainstream schools and contained about half as many items as the regular test instrument. These testing items were deemed more suitable for students with special education needs. A UH student questionnaire was also administered, which only contained trend items from the regular student questionnaire. The timing structure of both the UH test instrument and UH student questionnaire allowed more time per question than the regular instruments and UH sessions were generally held in small groups.

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 focused on science literacy (including collaborative problem solving and financial literacy as optional domains). In 2018, PISA focused on reading literacy and included global competence and financial literacy as optional domains.

Data Availability
Information on the availability of data and schedule of releases for PISA can be found at https://nces.ed.gov/surveys/pisa/.

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. This section uses 'jurisdictions' and 'education systems' interchangeably. Performance in each subject area can be compared across jurisdictions in terms of:

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

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 core 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.

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 consecutive days. Most education systems conducted testing from March through August 2018. The United States and the United Kingdom were given permission to move the testing dates to October through November in an effort to improve response rates. In the United States, students born between July 1, 2002, and June 30, 2003, were eligible to participate in PISA 2018.

The U.S. PISA 2018 national school sample consisted of 257 schools. This number represents an increase from the international minimum requirement of 150 and was implemented to offset anticipated 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 2015–16 Common Core of Data (CCD) and the data for private schools were from the 2015–16 Private School Universe Survey (PSS). Any school containing at least one of grades 7 through 12 was included in the school sampling frame. Participating schools provided a list of 15-year-old students (typically in August or September 2018) from which the sample was drawn using sampling software provided by the international contractor.

International Sample Design

The sample design for PISA 2018 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.

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). The target cluster size for countries/economies participating in the international option of financial literacy (FL) was increased to 52 students. A minimum of 6,300 students from a minimum of 150 schools was required in each country that planned to administer computer-based assessments. Education systems that opted to conduct paper-based assessments were required to assess a minimum of 5,250 students from a minimum of 150 schools. 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. For countries administering financial literacy, an additional sample of students was selected. If a jurisdiction had fewer than 5,250 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.

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, very small schools, and special education 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.

Response Rate Targets

School response rates. The PISA international guidelines for the 2018 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. Education systems 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. The international technical standards required a minimum participation rate of 80 percent of sampled students from schools (sampled and replacement) within each jurisdiction. This target applied in aggregate, not to each individual school. 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. Replacement students within a school were not allowed. 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 PISA 2018 school sample was drawn for the United States by the PISA consortium. The U.S. PISA sample was stratified into 8 explicit groups based on region of the country (Northeast, Central, West, Southeast) and control of school (public or private). Within each stratum, the frame was sorted for sampling by five categorical stratification

[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.

variables: grade range of the school (five categories); type of location relative to populous areas (city, suburb, town, rural); 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 U.S. PISA 2018 national school sample consisted of 257 schools, which was higher than the international sampling minimum of 150 to offset anticipated school nonresponse and ineligibility. The U.S. national sample included both public and private schools. Of the 52 students who were randomly sampled within each school, 41 students took the mathematics, science and reading literacy assessments and 11 students took the optional financial literacy assessment. The group of students who took the financial literacy assessment were referred to as the "financial literacy sample". Note that this was different from the approach used in the 2015 cycle, when financial literacy was administered to a subset of the students in the main PISA sample.

A total of 162 schools participated in the administration of U.S. national PISA, including 136 participating schools sampled as part of the original sample and 26 schools sampled as replacements for nonparticipating "original" schools. The overall weighted school response rate after replacements was 76 percent. For the United States as a whole, the weighted student response rate was 85 percent and the student exclusion rate was 4 percent.

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 an assessment like PISA, NCES requires that a nonresponse bias analysis be conducted when the response rate for schools falls below 85 percent or the response rate for students falls below 85 percent.

Assessment Design

Test scope and format. In PISA 2018, the three subject domains were tested, with reading literacy as the major domain and science and mathematics as the minor domains. Financial literacy was an optional domain administered by 21 education systems including the United States. An innovative (optional) domain in this cycle was global competence and the United States didn't participate in this domain.

The development of the PISA 2018 assessment instruments was an interactive process among the PISA Consortium, various expert committees, and OECD members. All mathematics and science items in the 2018 assessment instrument were trend items from previous assessments. Reading literacy and financial literacy items included both trend items and new items developed for 2018. 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 2018 was a computer-based assessment in most jurisdictions, including the United States. Test formats included multiple-choice and open response. Approximately 60 to 65 percent of items were multiple-choice and 35 to 40 percent were open response across reading, mathematics and science. Open response items were graded by trained scorers.

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.

Test design. PISA 2018 computer-based assessment was designed as a two-hour test. For the main subject of reading, material equivalent to 15 clusters of 30 minutes was developed. A multi-stage adaptive approach in reading assessment was adopted in PISA 2018. The reading material was organized into blocks instead of clusters. The reading assessment was composed of a core stage followed by stage 1 and stage 2. Students first saw a short Core stage, which consisted of between 7 and 10 items. The vast majority of these items (at least 80 percent and always at least 7 items) were automatically scored. Students' performance in this stage was provisionally classified as low, medium, or high, depending on the number of correct answers to these automatically scored items. The various Core Blocks of material delivered to students did not differ in any meaningful way in their difficulty. Stage 1 and 2, however, both existed in two different forms: comparatively easy and comparatively difficult. Students who displayed medium performance in the Core stage were equally likely to be assigned an easy or a difficult Stage 1. Students who displayed low performance in the Core stage had a 90 percent chance of being assigned to an easy Stage 1 and a 10 percent chance of being assigned to a difficult Stage 1. Students who displayed high performance in the Core stage had a 90 percent chance of being assigned to a difficult Stage 1 and a 10 percent chance of being assigned to an easy Stage 1. Students were assigned to easy and difficult Stage 2 blocks of material in much the same way.

In PISA 2018, in addition to the typical reading literacy items, the reading literacy assessment included a measure of reading fluency in the form of sentence processing. This measure required students to make a sensibility judgment about sentences of increasing complexity. Each student was assigned two fluency clusters for a total of 21 or 22 sentences before the reading literacy clusters. The reading fluency task were administered within a 3-minute timed session. Any sentences not completed within the three-minute session were skipped.

To measure trends in the subjects of mathematics and science, six clusters were included in each subject. In addition, four clusters of global competence items were developed. There was a total of 72 different test forms. Students spent one hour on the reading assessment plus one hour on one or two other subjects – mathematics, science or global competence. The financial literacy assessment lasted one hour (in addition to the regular PISA assessment) and comprised two clusters distributed to a subsample of students in combination with the reading and mathematics assessments.

For countries like the United States that took part in the core computer-based assessment (CBA) and the optional financial literacy assessment but did not opt to take part in the optional global competency domain, a total of 36 CBA testing forms were assembled for the assessment. 92 percent of students received forms numbered 1–24 of these 36 forms while 8 percent of students received forms numbered 25–36. These percentages are based on random assignment of test forms to students across schools.

Countries that used paper-based delivery for the main survey measured student performance with 30 pencil-and-paper forms containing trend items in the three core PISA subjects: reading, mathematics and science. Each form included one hour of reading items and items from at least one of the other two core domains. As a result, all students were administered two clusters of reading items, 46 percent of participating students were administered two clusters of mathematics items, 46 percent were administered two clusters of science items, and 8 percent were administered one cluster of mathematics and one cluster of science items, thus providing the covariance information about the three domains.

Data Collection and Processing

PISA 2018 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 2018, the U.S. national data collection contractor was Westat as well as a subcontractor, Pearson.

The 2018 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 2018 consisted of the Educational Testing Service (ETS), the U.S. research company Westat, cApStAn Linguistic Quality Control, Pearson, the German Institute for International Education Research (DIPF), Statistics Canada in Canada, University of Liège – aSPE in Belgium, University of Luxembourg in Luxembourg, and Australian Council for Educational Research (ACER) in Australia.

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. Most education systems conducted testing from March through August 2018. The United States and the United Kingdom were given permission to move the testing dates to September through December 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. The PISA consortium emphasized the use of standardized procedures in all education systems. Each economy collected its own data, based on detailed manuals provided by the PISA consortium (Westat 2014) that explained the survey's implementation, including precise instructions for the work of school coordinators and test administrators 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 each education system, a PISA Quality Monitor (PQM) who was engaged independently by the PISA consortium observed test administrations in a subsample of participating schools. The schools in which the independent observations were conducted were selected jointly by the PISA consortium and the PQM. In the United States, there were five PQMs who observed 15 schools from the national sample. 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.

Scoring. A substantial portion of the PISA 2018 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.

In PISA 2018, the process used for the main survey coding training was slightly different from that employed prior to the field trial as it included full training for all main survey items, both new and trend items. The coder query service was again used in the main survey as it had been in the field trial to assist countries in clarifying any uncertainty around the coding process or students' responses. Queries were reviewed, and responses were provided by domain-specific teams including item developers and members of the response team from previous cycles. Revisions were made to the coding guides for reading and global competence following the field trial and field trial pilot, respectively. The coder queries helped test developers see response categories that weren't anticipated during the development of the coding guide. Thus, based on the queries received, test developers made some coding guides clearer and added sample responses to the guides to better illustrate different types of responses.

Data entry and verification. In PISA 2018, a National Project Manager (NPM) in each jurisdiction was responsible for administering the assessments and collecting 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. The Student Delivery System (or SDS), a self-contained set of applications, was used to deliver the PISA 2018 CBA assessments and computer-based student background questionnaires. A master version was assembled first for countries to test within their national IT structure. This allowed countries to become familiar with the operation of the SDS and to check the compatibility of the software with computers being used to administer the assessment. After all components of national materials were locked, including the questionnaires and cognitive instruments, the student delivery system was assembled and tested first. The SDS was then released to countries for national testing. Countries were asked to check their SDS following a specific testing plan provided and to identify any residual content or layout issues. Where issues were identified, those were corrected and a second SDS was released. Once countries signed off on their national SDS, their instruments were released for the field trial and the main survey.

Harmonization or harmonizing variables is a process of mapping the national response categories of a particular variable into the international response categories so they can be compared and analyzed across countries. Not every nationally-adapted variable required harmonization, but for those that required harmonization, the Data Management team assisted the Background Questionnaire contractor with creating the harmonization mappings for each country with SAS code. This code was implemented into the data management cleaning and verification software in order to handle these harmonized variables during processing. ETS Data Management collaborated with the Background Questionnaire contractor to develop a series of validation checks that were performed on the data following harmonization.

Estimation Methods

Weighting. The use of sampling weights is necessary for computing 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 2018 was coordinated by Westat, as part of the international 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.

The student nonresponse adjustment was done within cells based first on their school nonresponse cell and their explicit stratum; within that, grade and gender were used when possible.

All PISA analyses were conducted using these adjusted sampling weights.

Scaling. For PISA 2018, item response theory (IRT) was used to estimate average scores for reading, science, and mathematics literacy for each education system, as well as for three reading process and three reading content subscales. For education systems participating in the financial literacy assessment these assessments were scaled separately and assigned separate scores. Scores for students were estimated as plausible values because each student completed only a subset of items. Ten plausible values were estimated for each student for each scale. These values represented 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, science, mathematics, and financial literacy scales are based on plausible values. In PISA, the reading, science, mathematics and financial literacy scales are from 0—1,000.

The PISA 2015 main study computer-based assessment included six clusters from each of the trend domains of science, reading, and mathematics literacy, six clusters of new science literacy test items, and three clusters of new collaborative problem-solving materials. The clusters were allocated in a rotated design to create six groups of test forms. Every student taking the assessment answered science items, and at least one but up to two of the other subjects of mathematics literacy, reading literacy, and/or collaborative problem solving. Students who were subsampled for the financial literacy assessment returned for a second session in which the focus was only on financial literacy and the accompanying student questionnaire.

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.

For PISA 2012, IRT was used 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 education systems participating in the financial literacy assessment and the computer-based assessment, these assessments were scaled separately and assigned separate scores.

For PISA 2015, IRT was used to estimate average scores for science, reading, and mathematics literacy for each economy, as well as for three science process and three science content subscales. For education systems participating in the financial literacy assessment and the collaborative problem-solving assessment, these assessments were 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. Ten plausible values were estimated for each student for each scale in PISA 2015 and PISA 2018. 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 ten relevant plausible value variables. The results of these ten analyses are averaged; then, significance tests that adjust for variation between the ten sets of results are computed.

Imputation. Missing background data from student and school questionnaires are not imputed for PISA 2009 reports. PISA 2015 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 science was assessed in 2000 and 2003, because the science framework was revised for 2006, it is possible to look at changes in science only from 2006 forward. Similarly, although reading was assessed in 2000, 2003, and 2006, and mathematics was assessed in 2000, because the reading framework was revised for PISA 2009 and the mathematics framework was revised for PISA 2003, it is possible to look at changes in reading only from 2009 forward and in mathematics only from 2003 forward. Although the PISA 2012 framework was updated, it is still possible to measure trends over time, as the underlying construct is intact. For specific trends in performance results, please see the NCES PISA website (https://nces.ed.gov/surveys/pisa/pisa2018/index.asp#/).

The PISA 2000, 2003, 2006, 2009, 2012, ,2015 and 2018 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. All mathematics and reading items in the PISA 2015 assessment instrument were trend items from previous assessments. Science items included both trend items and new items developed for 2015. In PISA 2018, all mathematics and science items were trend items and reading items included 238 new items and 72 trend items.

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 2018 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 2021.

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 2017–18 school year was just one of many possible samples that could have been selected. Therefore, estimates produced from the PISA 2018 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 2015, 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 2012, the United States reported 95 percent coverage of the national desired target population was achieved. For PISA 2015, the United States reported 83.5 percent coverage of the 15-year-old population and 96.7 coverage of national desired population. With a 3.3 percent overall exclusion rate, the United States reported a rate lower than the internationally acceptable exclusion rate of 5 percent. In PISA 2018, the United States reported 86.1 percent coverage of the 15-year old population and 96.2 percent coverage of national desired population with a 3.8 percent overall exclusion rate.

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 257 original sampled schools in the PISA 2018 United States national sample, 162 agreed to participate. The weighted school response rate before replacement was 65 percent for the United States, requiring NCES to conduct a nonresponse bias analysis, which was used by the PISA consortium and the OECD to evaluate the quality of the final United States sample.

Table PISA-1. U.S. weighted school and student response rates: PISA 2018

	Weighted response rate (percent)
School	
Before replacement	65.0
After replacement	76.4
Student	84.8

SOURCE: Organization for Economic Cooperation and Development (OECD), Program for International Student Assessment (PISA), 2018.

A total of 162 schools participated in the administration of national PISA, including 136 participating schools sampled as part of the original sample and 26 schools sampled as replacements for nonparticipating "original" schools. The overall weighted school response rate after replacements was 76.4 percent. For the United States as a whole, the weighted student response rate was 84.8 percent and the student exclusion rate was 3.8 percent.

For PISA 2015, 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 2015 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 2012–13 Common Core of Data for public schools and from the 2011–12 Private School Universe Survey for private schools. The available school characteristics included affiliation (public or private), locale (central city, suburb, town, rural), Census region, number of age-eligible students, total number of students, and percentage of various racial/ethnic groups (White, non-Hispanic; Black, non-Hispanic; Hispanic; Asian; American Indian or Alaska Native; Hawaiian/Pacific Islander; and two or more races). The percentage of students eligible for free or reduced-price lunch was available for public schools only.

For the United States original sample schools, schools in the Northeast were underrepresented among participating schools relative to eligible schools (12.6 vs. 17.1 percent, respectively), while schools in the South were overrepresented among participating schools (43.3 vs. 37.8 percent, respectively). Participating schools had a lower mean percentage of White, non-Hispanic students than the eligible sample (49.1 vs. 53.1 percent, respectively) and a higher mean percentage of Hispanic students than the eligible sample (27.4 vs. 24.6 percent, respectively). Additionally, the absolute value of the relative bias for private schools and schools in towns is greater than 10 percent, which indicates potential bias even though no statistically significant relationship was detected. When all factors were considered simultaneously in a logistic regression analysis, none of the parameter estimates were significant predictors 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 the United States final sample schools (with substitutes), there were no statistically significant relationships between participation status and any of the characteristics studied. However, the absolute value of the relative bias for private schools, schools in towns and the Northeast region are greater than 10 percent, which indicates potential bias even though no statistically significant relationships were detected. 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.

In the United States final sample schools with substitutes when school nonresponse adjusted weights were used for the participating schools, there were no statistically significant relationships between participation status and any of the characteristics studied. We therefore conclude that 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 2015 provides evidence that there is little potential for nonresponse bias in the PISA participating sample based on the characteristics studied. It also suggests that the use of substitute schools substantially reduced the potential for bias. Moreover, after the application of school nonresponse adjustments, there is no evidence of resulting potential bias in the final sample.

For PISA 2018, nonresponse bias analyses were again conducted at the school level in the U.S. sample as the weighted school response rate was below 85 percent. The general approach taken involved an analysis in three parts: (1) Analysis of the participating original sample: The distribution of the participating original school sample was compared with that of the total eligible original school sample. (2) Analysis of the participating final school sample with substitutes: The distribution of the participating final school sample, which included participating substitutes that were used as replacements for nonresponding schools from the eligible original sample, was compared to the total eligible final school sample. (3) Analysis of the nonresponse adjusted final sample with substitutes: The same sets of schools were compared as in the second analysis, but this time, when analyzing the participating final schools alone, school nonresponse adjustments were applied to the size-adjusted school base weights. The international weighting procedures form nonresponse adjustment classes by cross classifying the explicit and implicit stratification variables. The eligible sample were again weighted by their size-adjusted school base weights.

In addition to these tests, logistic regression models were used to provide a multivariate analysis that examined the conditional independence of these school characteristics as predictors of participation. The logistic regression compared frame characteristics for participating schools with non-participating schools. Multivariate analysis can provide additional insights, over and above those gained through the bivariate analysis. It may be the case that only one or two variables were actually related to participation status. However, if these variables were also related to the other variables examined in the analyses, then other variables, which were not related to participation status, would appear as significant in simple bivariate tables. Multivariate analysis, in contrast, examined the conditional relationships with participation after controlling for the other predictor variables—thereby, testing the robustness of the relationships between school characteristics and participation.

For original sample schools (not including substitute schools), nine variables were found to be statistically significantly related to participation in the bivariate analysis: school control, census region, poverty level, total school and age-eligible enrollments, White, non-Hispanic, Black, non-Hispanic, Hispanic, and free or reduced-price lunch. Additionally, the absolute value of the relative bias for small sized and large sized schools, American Indian or Alaska Native, and Hawaiian/Pacific Islander was greater than 10 percent, which indicated potential bias even though no statistically significant relationship was detected.

For the final sample of schools (with substitute schools) with school nonresponse adjustments applied to the weights, no variables were found to be statistically significantly related to participation in the bivariate analysis. However, the absolute value of the relative bias for small sized schools and Hawaiian/Pacific Islander was greater than 10 percent.

In sum, the investigation into nonresponse bias at the school level in the U.S. PISA 2018 data provides evidence that there is some potential for nonresponse bias in the PISA participating original sample based on the characteristics studied. It also suggests that, while there is some evidence that the use of substitute schools reduced the potential for bias, it has not reduced it substantially. However, after the application of school nonresponse adjustments, there is little evidence of resulting potential bias in the available frame variables and correlated variables 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 in 2012 and 2014, 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 on PISA, contact:

Samantha Burg
Phone: (202) 245-7537
E-mail: Samantha.Burg@ed.gov

Mailing Address
National Center for Education Statistics
Institute of Education Sciences
Potomac Center Plaza
550 12th Street, SW
Washington, DC 20202

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. https://nces.ed.gov/pubs2008/2008016.pdf.

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. https://nces.ed.gov/pubs2011/2011004.pdf.

Kastberg, D., Perkins, R., Cummings, L., Ferraro, D., and Goodnow, M. (Forthcoming). *Technical Report and User Guide for the 2018 Program for International Student Assessment (PISA)*. (Forthcoming). U.S. Department of Education. Washington, D.C.: National Center for Education Statistics.

Kastberg, D., Chan, J.Y., and Murray, G. (2016). *Performance of U.S. 15-Year-Old Students in Science, Reading, and Mathematics Literacy in an International Context: First Look at PISA 2015* (NCES 2017-048). National Center for Education Statistics, Institute of

Education Sciences, U.S. Department of Education. Washington, DC. https://nces.ed.gov/pubs2017/2017048.pdf.

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. https://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. https://nces.ed.gov/surveys/pisa/pdf/2011025.pdf.

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. https://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. https://nces.ed.gov/pubs2005/2005003.pdf.

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.

Organization for Economic Cooperation and Development (OECD). (2017). *PISA 2015 Technical Report*. Paris: Author.

Organization for Economic Cooperation and Development (OECD). (2020). *PISA 2018 Technical Report.* Paris: Author.

Sen, A., Mai, T., Kastberg, D., Cummings, L., and Perkins, R. (2019). *Highlights of U.S. PISA 2018 Results Web Report* (NCES 2020-166 and 2020-072). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC. https://nces.ed.gov/surveys/pisa/pisa2018/index.asp#/.

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. https://nces.ed.gov/surveys/pisa/pdf/2009055.pdf.

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. https://www.oecd.org/pisa/pisaproducts/PISA%202012%20framework%20e-book_final.pdf.

Organization for Economic Cooperation and Development (OECD). (2014), *PISA 2012 Technical Report.*

http://www.oecd.org/pisa/pisaproducts/pisa2012technicalreport.htm.

PISA Project Consortium. (2005b). *School Sampling Preparation Manual: PISA 2006 Main Study.* Originally retrieved on (November 14, 2008) from a link that is currently inactive. Retrieved on (October 23, 2020) from (http://www.oecd.org/education/school/programmeforinternationalstudentassessmentpisa/39829698.pdf)

PISA Project Consortium. (2005c). *PISA 2006 Main Study Test Administrator's Manual.* Originally retrieved on (November 14, 2008) from a link that is currently inactive. Retrieved on (October 23, 2020) from (https://www.acer.org/files/pisa2006_test_administrator_manual.pdf)

PISA Project Consortium. (2005d). *PISA 2006 Main Study School Coordinator's Manual.* Originally retrieved on (November 14, 2008) from a link that is currently inactive. Retrieved on (October 23, 2020) from (https://www.acer.org/files/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. https://nces.ed.gov/pubs2007/2007049.pdf.

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. https://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. https://nces.ed.gov/pubs2001/200107.pdf.

Stephens, M., and Coleman, M. (2007). *Comparing PIRLS and PISA With NAEP in Reading, Mathematics, and Science*. National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC. Originally retrieved on (November 14, 2008) from a link that is currently inactive. Retrieved on (October 23, 2020) from (https://nces.ed.gov/surveys/pisa/pdf/comppaper12082004.pdf)