APPENDIX

Methodology

Table of Contents

Introduction

Science and Engineering Indicators (*Indicators*) contains data compiled from a variety of sources. This appendix explains the methodological and statistical criteria used to assess possible data sources for inclusion in *Indicators* and to develop statements about the data. It also provides basic information about how statistical procedures and reasoning are applied.

This appendix has four main sections, a glossary, and information on viewing the data sources for this report. The first section describes the considerations that are part of the selection process for information to be included in *Indicators*. The second section discusses the different sources of information (e.g., sample surveys, censuses, and administrative records) used in the report and provides details about each type. The third section discusses factors that can affect accuracy at all stages of the survey process. The fourth section discusses the statistical testing used to determine whether differences between sample survey-based estimates are *statistically significant*—that is, greater than could be expected by chance. The glossary covers statistical terms commonly used or referred to in the text. The appendix concludes by providing information on how to access the report's data sources, which can be viewed by chapter and by data provider.

Selection of Data Sources

Information is available from many sources, and it can vary in substantial ways. Several criteria guide the selection of data for *Indicators*:

Representativeness. Data should represent the entire national or international populations of interest and should reflect the heterogeneity of those populations. Data should be also available for the subdomains of interest covered in *Indicators* (e.g., the population of scientists and engineers or the topic of R&D spending by universities).

Relevance. Data should include indicators central to the functioning of the science and technology enterprise.

Timeliness. Data that are not part of a time series should be timely (i.e., they should be the most recent data available that meet the selection criteria).

Statistical and methodological quality. Survey methods used to collect data should provide sufficient assurance that survey estimates are robust and that statements based on statistical analysis of the data are valid and reliable. Nonsurvey data, such as administrative records, or data from other third-party sources should similarly be assessed for quality—that is, fitness for use. All external data should be properly sourced and cited. Data included in *Indicators* must be of high quality. Known limitations of the external data must be clearly stated. Data quality has several characteristics. Some key dimensions of quality include the following.

Validity. Data have *validity* if they accurately measure the phenomenon they are supposed to represent.

Reliability. Data have *reliability* if similar results would be produced if the same measurement or procedure were performed multiple times on the same population.

Accuracy. Data are *accurate* if estimates from the data do not widely deviate from the true population value.

Data that are collected by U.S. government agencies and that are products of the federal statistical system meet the rigorous statistical and methodological criteria described above. Unless otherwise indicated, these data are representative of the nation as a whole and of the demographic, organizational, or geographic subgroups that constitute it.

For data collected by governments in other countries and by nongovernment sources, including private survey firms and academic researchers, methodological information is examined to assess conformity with the criteria that U.S. federal agencies typically use. Government statistical agencies in the developed world cooperate extensively both in developing data-quality standards and in improving international comparability for key data, and these agencies ensure that the methodological information about the data generated by this international statistical system is relatively complete.

Often, methodological information about data from nongovernmental sources and from governmental agencies outside the international statistical system is less well documented. These data must meet basic scientific standards for representative sampling of survey respondents and for adequate and unbiased coverage of the population under study. The resulting measurements must be sufficiently relevant and meaningful to warrant publication despite methodological uncertainties that remain after the documentation has been scrutinized.

Many data sources that contain pertinent information about a segment of the S&E enterprise are not cited in *Indicators* because their coverage of the United States is partial in terms of geography, incomplete in terms of segments of the population, or otherwise not representative. For example, data may be available for only a limited number of states, or studies may be based on populations not representative of the United States as a whole. Similarly, data for other countries should cover and be representative of the entire country. In some cases, data that have limited coverage or are otherwise insufficiently representative are referenced in sidebars.

Types of Data Sources

Much of the data cited in *Indicators* comes from surveys. Surveys strive to measure characteristics of target populations. To generalize survey results correctly to the population of interest, a survey's *target population* must be rigorously defined, and the criteria determining membership in the population must be applied consistently in determining which units to include in the survey. After a survey's target population has been defined, the next step is to establish a list of all members of that target population (i.e., a *sampling frame*). Members of the population must be selected from this list using accepted statistical methods so that it will be possible to generalize from the sample to the population as a whole. Surveys sometimes sample from lists that, to varying extents, omit members of the target population because complete lists are typically unavailable.

Some surveys are censuses (also known as *universe surveys*), in which the survey attempts to obtain data for all population units. The decennial census, in which the target population is all U.S. residents, is the most familiar census survey. *Indicators* uses data from the Survey of Earned Doctorates, an annual census of individuals who earn research doctorates from accredited U.S. institutions, for information about the numbers and characteristics of new U.S. doctorate holders.

Other surveys are *sample surveys*, in which data are obtained for only a portion of the population units. Samples can be drawn using either probability-based or nonprobability-based sampling procedures. A sample is a *probability sample* if each unit in the sampling frame has a known, nonzero probability of being selected for the sample. Probability samples are preferred because their use allows the computation of measures of precision and the subsequent statistical evaluation of inferences about the survey population. An example of a sample survey is the National Survey of College Graduates (NSCG). The NSCG gathers data on the nation's college graduates, with particular focus on those educated or employed in an S&E field. In *nonprobability sampling*, the sample is drawn with an unknown probability of selection. Polls that elicit responses from self-selected individuals, such as opt-in Internet surveys or phone-in polls, are examples of nonprobability sample surveys. Except for some Asian surveys referenced in Chapter 7, sample surveys included in *Indicators* use probability sampling.

Surveys may be conducted of individuals or of organizations, such as businesses, universities, or government agencies. Surveys of individuals are referred to as *demographic surveys*. Surveys of organizations are often referred to as *establishment surveys.* An example of an establishment survey used in *Indicators* is the Higher Education Research and Development Survey.

Surveys may be longitudinal or cross-sectional. In a *longitudinal survey*, the same sample members are surveyed repeatedly over time. The primary purpose of longitudinal surveys is to investigate changes over time. The Survey of Doctorate Recipients is a sample survey of individuals who received research doctorates from U.S. institutions. The survey was originally designed to produce cross-sectional estimates, but the data have also been adapted by researchers to conduct longitudinal studies. *Indicators* uses results from this survey to analyze the careers of doctorate holders.

Cross-sectional surveys provide a snapshot at a given point in time. When conducted periodically, cross-sectional surveys produce repeated snapshots of a population, also enabling analysis of how the population changes over time. However, because the same individuals or organizations are not included in each survey cycle, cross-sectional surveys cannot, in general, track changes for specific individuals or organizations. National and international assessments of student achievement in K–12 education, such as those discussed in Chapter 1, are examples of repeated cross-sectional surveys. Most of the surveys cited in *Indicators* are conducted periodically, although the frequency with which they are conducted varies.

Surveys can be self- or interviewer-administered, and they can be conducted using a variety of modes (e.g., postal mail, telephone, the Web, e-mail, or in person). Many surveys are conducted using more than one mode. The NSCG is an example of a multimode survey. It is conducted primarily via the Web; potential participants who do not respond to the questionnaire are contacted via telephone.

Some of the data in *Indicators* come from *administrative records* (data collected for the purpose of administering various programs). Examples of data drawn directly from administrative records in *Indicators* include patent data from the records of government patent offices; bibliometric data on publications in S&E journals, compiled from information collected and published by the journals themselves; and data on foreign S&E workers temporarily in the United States, drawn from the administrative records of immigration agencies.

Many of the establishment surveys that *Indicators* uses depend heavily, although indirectly, on administrative records. Universities and corporations that respond to surveys about their R&D activities often use administrative records developed for internal management or income tax reporting purposes to respond to these surveys.

Data Accuracy

Accurate information is a primary goal of censuses and sample surveys. Accuracy can be defined as the extent to which results deviate from the true values of the characteristics in the target population. Statisticians use the term “error” to refer to this deviation. Good survey design seeks to minimize survey error.

Statisticians usually classify the factors affecting the accuracy of survey data into two categories: nonsampling and sampling errors. *Nonsampling error* applies to administrative records and surveys, including censuses, whereas *sampling error* applies only to sample surveys.

Nonsampling Error

Nonsampling error refers to error related to the design, data collection, and processing procedures. Nonsampling error may occur at each stage of the survey process and is often difficult to measure. The sources of nonsampling error in surveys have analogues for administrative records: the purposes for and the processes through which the records are created affect how well the records capture the concepts of interest of relevant populations (e.g., patents, journal articles, immigrant scientists and engineers). A brief description of five sources of nonsampling error follows. For convenience, the descriptions refer to samples, but they also apply to censuses and administrative records.

Specification error. Survey questions often do not perfectly measure the concept for which they are intended as indicators. For example, the number of patents does not perfectly quantify the amount of invention.

Coverage error. The sampling frame, the listing of the target population members used for selecting survey respondents, may be inaccurate or incomplete. If the frame has omissions, duplications, or other flaws, the survey is less representative because coverage of the target population is inaccurate. Frame errors often require extensive effort to correct.

Nonresponse error. Nonresponse error can occur if not all members of the sample respond to the survey. *Response rates* indicate the proportion of sample members that respond to the survey. Response rate is not always an indication of nonresponse error.

Nonresponse can cause *nonresponse bias*, which occurs when the people or establishments that respond to a question, or to the survey as a whole, differ in systematic ways from those who do not respond. For example, in surveys of national populations, complete or partial nonresponse is often more likely among lower-income or less-educated respondents. Evidence of nonresponse bias is an important factor in decisions about whether survey data should be included in *Indicators*.

Managers of high-quality surveys, such as those in the U.S. federal statistical system, do research on nonresponse patterns to assess whether and how nonresponse might bias survey estimates. *Indicators* notes instances where reported data may be subject to substantial nonresponse bias.

Measurement error. There are many sources of measurement error, but respondents, interviewers, mode of administration, and survey questionnaires are the most common. Knowingly or unintentionally, respondents may provide incorrect information. Interviewers may influence respondents' answers or record their answers incorrectly. The questionnaire can be a source of error if there are ambiguous, poorly worded, or confusing questions, instructions, or terms or if the questionnaire layout is confusing.

In addition, the records or systems of information that a respondent may refer to, the mode of data collection, and the setting for the survey administration may contribute to measurement error. Perceptions about whether data will be treated as confidential may affect the accuracy of survey responses to sensitive questions, such as those about business profits or personal incomes.

Processing error. Processing errors include errors in recording, checking, coding, and preparing survey data to make them ready for analysis.

Sampling Error

Sampling error is the most commonly reported measure of a survey's precision. Unlike nonsampling error, sampling error can be quantitatively estimated in most scientific sample surveys.

Sampling error is the uncertainty in an estimate that results because not all units in the population are measured. Chance is involved in selecting the members of a sample. If the same, random procedures were used repeatedly to select samples from the population, numerous samples would be selected, each containing different members of the population with different characteristics. Each sample would produce different population estimates. When there is great variation among the samples drawn from a given population, the sampling error is high, and there is a large chance that the survey estimate is far from the true population value. In a census, because the entire population is surveyed, there is no sampling error, but nonsampling errors may still exist.

Sampling error is reduced when samples are large, and most of the surveys used in *Indicators* have large samples. Typically, sampling error is a function of the sample design and size, the variability of the measure of interest, and the methods used to produce estimates from the sample data.

Sampling error associated with an estimate is often measured by the coefficient of variation or margin of error, both of which are measures of the amount of uncertainty in the estimate.

Statistical Testing of Sample Survey Data

Statistical tests can be used to determine whether differences observed in sample survey data are “real” differences in the population. Differences that are termed *statistically significant* are likely to occur in the target population. When *Indicators* reports statements about differences on the basis of sample surveys, the differences are statistically significant at least at the 10% level. This means that, if there were no true difference in the population, the chance of drawing a sample with the observed or greater difference would be no more than 10%.

A statistically significant difference is not necessarily large, important, or significant in the usual sense of the word. It is simply a difference that is unlikely to be caused by chance variation in sampling. With the large samples common in *Indicators* data, extremely small differences can be found to be statistically significant. Conversely, quite large differences may not be statistically significant if the sample or population sizes of the groups being compared are small. Occasionally, apparently large differences are noted in the text as not being statistically significant to alert the reader that these differences may have occurred by chance.

Numerous differences are apparent in every table in *Indicators* that reports sample data. The tables permit comparisons between different groups in the survey population and in the same population in different years. It would be impractical to test and indicate the statistical significance of all possible comparisons in tables involving sample data.

As explained in the section About Science and Engineering Indicators, *Indicators* presents indicators. It does not model the dynamics of the S&E enterprise, although analysts could construct models using the data in *Indicators*. Accordingly, *Indicators* does not make use of statistical procedures suitable for causal modeling and does not compute effect sizes for models that might be constructed using these data.

Glossary

Most glossary definitions are based on U.S. Office of Management and Budget, Office of Statistical Policy (2006), *Standards and Guidelines for Statistical Surveys* (https://unstats.un.org/unsd/dnss/docs-nqaf/USA_standards_stat_surveys.pdf), and U.S. Census Bureau (2006), *Organization of Metadata, Census Bureau Standard Definitions for Surveys and Census Metadata*. In some cases, glossary definitions are somewhat more technical and precise than those in the text, where fine distinctions are omitted to improve readability.

Accuracy: Accuracy is the difference between the estimate and the true parameter value.

Administrative records: Microdata records collected for the purpose of carrying out various programs (e.g., tax collection). Unlike survey data, administrative data were not originally collected for statistical purposes.

Bias: Systematic deviation of the survey estimated value from the true population value. Bias refers to systematic errors that can occur with any survey under a specific design.

Census: A data collection that seeks to obtain data directly from all eligible units in the entire target population. It can be considered a sample with a 100% sampling rate. A census may use data from administrative records for some units rather than direct data collection.

Coverage: Extent to which all elements on a frame list are members of the population and to which every element in a population appears on the frame list once and only once.

Coverage error: Discrepancy between statistics calculated on the frame population and the same statistics calculated on the target population. *Undercoverage* errors occur when target population units are missed during frame construction, and *overcoverage* errors occur when units are duplicated, enumerated incorrectly, or are not part of the target population.

Cross-sectional sample survey: Based on a representative sample of respondents drawn from a population at a particular point in time.

Estimate: A numerical value for a population parameter derived from information collected from a survey or other sources.

Estimation error: Difference between a survey estimate and the true value of the parameter in the target population.

Frame: A mapping of the universe elements (i.e., sampling units) onto a finite list (e.g., the population of schools on the day of the survey).

Item nonresponse: Occurs when a respondent fails to respond to one or more relevant items on a survey.

Longitudinal sample survey: Follows the experiences and outcomes over time of a representative sample of respondents (i.e., a cohort).

Measurement error: Difference between observed values of a variable recorded under similar conditions and some fixed true value (e.g., errors in reporting, reading, calculating, or recording a numerical value).

Nonresponse bias: Occurs when the observed value deviates from the population parameter due to systematic differences between respondents and nonrespondents. Nonresponse bias may occur as a result of not obtaining 100% response from the selected units.

Nonresponse error: Overall error observed in estimates caused by differences between respondents and nonrespondents. It consists of a variance component and nonresponse bias.

Nonsampling error: Includes specification errors and measurement errors due to interviewers, respondents, instruments, and mode; nonresponse error; coverage error; and processing error.

Parameter: An unknown, quantitative measure (e.g., total revenue, mean revenue, total yield, or number of unemployed people) for the entire population or for a specified domain of interest.

Population: The set of persons or organizations to be studied, which may not be of finite size.

Precision of survey results: How closely results from a sample can reproduce the results that would be obtained from a complete count (i.e., census) conducted using the same techniques. The difference between a sample result and the result from a complete census taken under the same conditions is an indication of the precision of the sample result.

Probabilistic methods: Any of a variety of methods for survey sampling that gives a known, nonzero probability of selection to each member of a target population. The advantage of probabilistic sampling methods is that sampling error can be calculated. Such methods include random sampling, systematic sampling, and stratified sampling. They do not include convenience sampling, judgment sampling, quota sampling, and snowball sampling.

Reliability: Degree to which a measurement technique would yield the same result each time it is applied. A measurement can be both reliable and inaccurate.

Response bias: Deviation of the survey estimate from the true population value due to measurement error from the data collection. Potential sources of response bias include the respondent, the instrument, the mode of data collection, and the interviewer.

Response rates: These measure the proportion of the sample frame represented by the responding units in each study.

Sample design: Refers to the combined target population, frame, sample size, and sample selection methods.

Sample survey: A data collection that obtains data from a sample of the frame population.

Sampling error: Error that occurs because all members of the frame population are not measured. It is associated with the variation in samples drawn from the same frame population. The sampling error equals the square root of the variance.

Standard error: Standard deviation of the sampling distribution of a statistic. Although the standard error is used to estimate sampling error, it includes some nonsampling error.

Statistical significance: Attained when a statistical procedure applied to a set of observations yields a *p*-value that exceeds the level of probability at which it is agreed that the null hypothesis will be rejected.

Target population: Any group of potential sample units or individuals, businesses, or other entities of interest.

Unit nonresponse: Occurs when a respondent fails to respond to all required response items (i.e., fails to complete or return a data collection instrument).

Universe survey: Involves the collection of data covering all known units in a population (i.e., a census).

Validity: Degree to which an estimate is likely to be true and free of bias (systematic errors).

View Data Sources

The complete list of data sources used in this volume can be found in the Data Sources section (https://www.nsf.gov/statistics/2018/nsb20181/data/sources). Data sources can be viewed by chapter and by data provider.