APPENDIX A.
TECHNICAL NOTES

APPENDIX A. TECHNICAL NOTES

The data on doctoral scientists and engineers contained in this report come from the 1993 Survey of Doctorate Recipients (SDR). The SDR has been conducted biennially since 1973 by the National Research Council (NRC) for the National Science Foundation (NSF). Additional data on education and demographic information come from the National Research Council's Doctorate Records File (DRF). The DRF contains data from an ongoing census of research doctorates earned in the United States since 1920.

In 1993, as part of a large redesign of the NSF surveys of scientists and engineers, the SDR underwent significant revisions to improve data quality and relevance to policy and research interests:

- The survey instrument was expanded and redesigned; questions were retooled to improve validity and comparability with data from other NSF and federal surveys.
- While continuing intensive response followup efforts begun in 1991, the 1993 SDR restored the 30-percent sample loss of the previous survey round.
- Imputation was introduced to compensate for item nonresponse.

These changes, in particular the extensive instrument redesign, will affect comparability with SDR data from earlier years. The user should be cautious when using 1993 data in time-series or longitudinal analyses.

THE SAMPLING FRAME AND TARGET POPULATION

For the 1993 SDR the sampling frame for scientists and engineers was selected from the DRF to include individuals who

(1) had earned a doctoral degree from a U.S. college or university in a science or engineering field; and

(2) were U.S. citizens or, if non-U.S. citizens, indicated they had plans to remain in the United States after degree award; and

(3) were under 76 years of age as of April 1993 (the survey reference date).

The 1993 frame consisted of graduates who had earned their degrees between January 1942 and June 1992. Persons who did not meet the age criteria (or had died) were eliminated from the sample.

The survey has two additional eligibility criteria for the survey target population. The sampled member must be resident in the United States and not institutionalized as of the reference date.

SAMPLE DESIGN

In 1993, the SDR sample size was 49,228. This represented an increase of 30 percent over the 1991 survey. The total sample was selected from 3 groups:

(1) 1991 sample members who were still eligible in 1993,

(2) most of the 1989 sample members who had been cut from the 1991 sample, and

(3) a sample of the 1991-92 graduating cohort.

Group 1 cases were included with certainty because they are the core sample that is conveyed from year to year; groups 2 and 3 cases were sampled and added to the core sample to form the total sample.

The basic sample design was a stratified random sample. The variables used for stratification were 15 broad fields of degree, 2 genders, and an 8-category "group" variable combining race/ethnicity, handicap status, and citizenship status.

The overall sampling rate was about 1 in 11 (9 percent) in the 1993 SDR, applied to a population of 568,700. However sampling rates varied considerably within and between the strata. These differences resulted from oversampling of women, minority groups and other groups of special interest, and the accumulation of sample size adjustments over the years.

Data Collection

In 1993, there were 2 phases of data collection: a mail survey and telephone followup interviewing with nonrespondents. The mail survey consisted of an advance letter and 2 waves of a personalized mailing package, with a reminder postcard between waves 1 and 2. The first-wave mailing was sent in May 1993, with the followup mailing in June. As part of an experiment to test the effectiveness of Priority Mail, about one-third of the sample were sent a second followup mailing in July.

Phase 2 consisted of telephone interviewing. All nonrespondents to the mail survey were followed up using computer-assisted telephone interviewing (CATI). Telephone interviewing was conducted between September 1993 and February 1994.

Survey Instrument Design and Content

In 1993, the SDR survey content and instrument went through a major redesign. The survey instrument, i.e., the wording and structure of the questions, changed greatly between 1991 and 1993. The format and layout of the questionnaires were changed to a more "respondent friendly" design to improve data quality. This included using a larger type size for improved readability, using graphical aids to indicate skip patterns, and using reverse printing to indicate answer spaces. The survey instrument was expanded from eight pages to twenty pages.

The survey content was also enhanced in 1993. These enhancements included

- the addition of new questions to gather information on such topics as degrees earned since receipt of the first doctorate, relationship of degree to current job, and reasons for making job changes;
- the expansion and modification of the sections on current employment and demographic characteristics to improve quality and validity; and
- replacing the concept of "employment field" to "occupation," allowing the analysis of the relationship of education to outcomes (occupation).

The reference period was changed to the "week of April 15th" from "September" in 1991 and "February" in earlier years. Thus, between the 1991 and 1993 surveys about 20 months had elapsed, as opposed to 32 months between the 1989 and 1991 surveys, and 24 months in predecessor years.

Response Rates

The overall response rate for the 1993 SDR was 88 percent. The response to the mail phase of the survey was about 61 percent. (Response rates were calculated as the weighted response divided by the in-scope sample cases.) Of the nonrespondents in the survey, it is estimated that about 40 percent were refusals, 35 percent were located but not interviewed, and 25 percent were not located.

Data Preparation

As completed survey mail questionnaires were received, they were logged and transferred to the editing and coding unit at the NRC for processing. The coders carried out a variety of checks to prepare the documents for data entry. Specifically, they resolved incomplete or contradictory answers, imputed missing answers if logically appropriate, reviewed "other specify" responses for possible backcoding to a listed response, and assigned numeric codes to open-ended questions such as employer name.

Once questionnaires were edited and coded, they were sent to data entry. The data entry program contained a full complement of range and consistency checks to check for entry errors and inconsistent answers. The range and consistency checks were also applied to the CATI data via batch processing. Further computer checks were performed to test for inconsistent values; these were corrected and the process repeated until no inconsistencies remained.

At this point, the survey data file was ready for imputation of missing data. As a first step, basic frequency distributions were produced to show

nonresponse rates to each question—these were generally less than 2 percent, with the exception of salary, which was 5.8 percent. Two methods for imputation were adopted. The first, cold decking, was used mainly for demographic variables that are static, i.e., not subject to change. Using this method, historical data provided by respondents in previous years were used to fill a missing response. For example, if a respondent indicated he was Asian in 1991, but left the item blank in 1993, then "Asian" was assigned to his race in 1993. In cases where no historical data were available, and for nondemographic variables (such as employment status, primary work activity, and salary), hot decking was used. This is the process of finding a donor with characteristics similar to the case with the missing value and using the response given by the donor as a proxy response. Hot decking involves creating groups of cases with common characteristics (through the cross-classification of auxiliary variables) and then selecting a donor at random for the case with the missing value. As a general rule, no data value was imputed from a donor in one cell to a recipient in another cell.

For a few variables, such as employer name and zip code, imputation was not performed.

Weighting and Estimation

The next phase of the survey process involved weighting the survey data to compensate for unequal probabilities of selection to the sample and to adjust for the effects of unit nonresponse. The first step was the construction of sampling weights, which were calculated as the inverse of the probability of selection, taking into account all stages of the sample selection process over time. The sampling weight can be viewed as the number of population members the sample member represents. Sampling weights varied within cells because different sampling rates were used depending on the year of selection and the stratification in effect at that time.

The next step was to adjust the sampling weights for unit nonresponse. (Unit nonresponse occurs when the sample member refuses to participate or cannot be located.) This was done in a group of nonresponse adjustment cells created using poststratification. Within each nonresponse adjustment cell, a weighted nonresponse rate, which took into account both mail and CATI nonresponse, was calculated. The nonresponse adjustment factor was the inverse of this weighted response rate. The initial set of nonresponse adjustment factors was examined and, under certain conditions, some of the cells were collapsed if use of the adjustment factor would create excessive variance.

The final weights for respondents were calculated by multiplying their respective sample weights by the nonresponse adjustment factor. In data analysis, population estimates are made by summing the final weights of all respondents who possess a particular characteristic.

Reliability[2]

The statistics in this report are subject to both sampling and nonsampling error. Sampling variability occurs because a sample rather than an entire population is surveyed. Sampling errors were developed using a generalized variance procedure in order to provide approximate sampling errors that would be applicable to a wide variety of items. As a result, these sampling errors provide an indication of the order of magnitude of a sampling error rather than a precise sampling error for any specific item.

Information provided in table A-3 permits the user to calculate approximate standard errors. The general form of the equation used to model the generalized variances is $V = a + b/x$, where V was modeled in relative standard error form.

The following computational form can be used for estimating the standard error of totals using the formula

$$S_x=[ax^2+bx]^{1/2}$$

where "x" equals the estimated total and "a" and "b" are the regression coefficients provided.

[2] The data and material on sampling reliability presented here are from The Methodological Report of the 1993 Survey of Doctorate Recipients (Washington, D.C. Office of Scientific and Engineering Personnel, National Research Council, forthcoming).

Values of "a" and "b" by S&E fields for selected groups are given in table A-3.[3]

Tables A-4 through A-8 present approximate standard errors associated with totals for different segments of the doctoral population. Tables A-9 through A-13 present standard error estimates for the estimated percent [4] of a subgroup having a particular characteristic.

The approximate standard error of percentages also was developed using the same general model form. Standard errors for percentages may be estimated using the computational formula

$$S_p = p[b((1/x)-(1/y))]^{1/2}$$

where p equals the percentage possessing the specific characteristic and x and y represent the numerator and denominator, respectively, of the ratio that yields the observed percentage.

In addition to sampling error, data are subject to nonsampling error. Sources of nonsampling error include nonresponse bias, which arises when individuals who do not respond to a survey differ significantly from those who do, and measurement error, which arises when we are not able to precisely measure the variables of interest. These sources of error are much harder to estimate than sampling errors.

NOTES ON THE TABLES

The following notes facilitate use of data in the detailed tables.

Because of the changes (described above) introduced to the 1993 SDR, users are advised that data in this report are not comparable with SDR data published by NSF for prior survey years.

Field of doctorate is the field of degree as specified by the respondent in the Survey of Earned Doctorates at the time of degree conferral.

Occupation data were derived from responses to several questions on the kind of work done by the respondent. The occupational classification of the respondent was based on his or her principal job held during the reference week—or last job held, if not employed on the reference week (questions A17 and A5). Also used in the occupational classification was a respondent-selected job code (questions A18 and A6).

Sector of employment was based on responses to questions A13 and A16. The category "universities and 4-year colleges" includes 4-year colleges or universities, medical schools (including university-affiliated hospitals or medical centers), and university affiliated research institutions. "Private-for-Profit" includes self-employed in incorporated business.

Geographic division was based primarily on responses to question A10 on the location of employment. Individuals not reporting place of employment were classified by their mailing address.

Place Of Birth categories were defined as follows:

U.S.	=	Fifty states plus the Virgin Islands, Panama Canal Zone, Puerto Rico, American Samoa, Trust Territory, and Guam
Latin	=	Mexico, Central America, Cuba and America Islands
South America	=	Argentina, Bolivia, Brazil, Chile, Columbia, Ecuador, French Guiana, Guyana, Paraguay, Peru, Suriname, Uruguay, Venezuela
Northern Europe	=	Denmark, England, Finland, Iceland, Northern Ireland,

[3] The generalized error estimates in this report were based on a set of assumptions that did not appear to hold in the case of some small subpopulations. In such cases, the parameters listed for a higher-level field within a demographic group or a higher-level demographic group within a field were considered a useful substitute as a generalized error estimate.

[4] The estimated percent is based on the ratio of two estimated totals, where the numerator is a subset of the denominator.

		Republic of Ireland, Norway, Scotland, Sweden, Wales
Central Europe	=	Austria, Germany, Italy, Liechtenstein, Malta
Western Europe	=	Andorra, Belgium, France, Gibraltar, Luxembourg, Monaco, The Netherlands, Portugal, Spain, Switzerland
Eastern Europe	=	Albania, Armenia, Azerbaijan, Belarus, Bosnia-Herzegovina, Bulgaria, Czech Republic, Croatia, Estonia, Georgia, Greece, Hungary, Kazakhstan, Kyrgyzstan, Latvia, Lithuania, Macedonia, Moldova, Poland, Romania, Russia, Slovakia, Slovenia, Tajikstan, Turkemnistan, Ukraine, Uzbekistan, Federal Republic of Yugoslavia
Eastern Asia	=	Cambodia, People's Republic of China, Taiwan, China Unspeci fied, Hong Kong, Japan, Republic of Korea, Korea Unspecified, Laos, Macao, Malaysia, Myanmar, Singapore, Thailand, Democratic Republic of Vietnam, Republic of Vietnam
Western	=	Afghanistan, Bahrain, Bangladesh, Asia Cyprus, India, Iran, Iraq, Israel, Jordan, Kuwait, Lebanon, Nepal, Pakistan pre-1971, Palestine, Saudi Arabia, Sri Lanka, Syria, Turkey
Australasia	=	Australia, Indonesia, New Zealand, Philippines
Africa	=	Algeria, Egypt, Ethopia, Ghana, Kenya, Libya, Morocco, Nigeria, South Africa, Sudan, Africa, not specified

Primary work activity was determined from responses to question A25. "Development" includes the development of equipment, products, and systems. "Design" includes the design of equipment, processes, and models.

Federal support was determined from responses to questions A31 and A32. The reference period used for these questions was changed in 1993. The 1993 questionnaire used "the week of" as the reference period whereas the 1991 questionnaire used "the past year."

Tenure status was obtained from the responses to question A15.

Salary data were derived from responses to question A29, in which information was requested regarding annual salary before deductions for income tax., social security, retirement, but excluding bonuses, overtime, and summer teaching. Salaries reported are median annual salaries, rounded to the nearest $100 and computed for full-time employed scientists and engineers only, excluding self-employed. For individuals employed by educational institutions, no accommodation was made to convert academic-year salaries to calendar-year salaries as in previous years. Prior to 1993, academic-year (9 to 10 months) salaries were multiplied by eleven-ninths to adjust to a calendar-year (11 to 12 months) scale.

Racial/ethnic data were based on responses to questions E5, E6, and E7. Individuals included in the Hispanic category are not included in other race/ethnicity categories as in previous years.

Selected Employment Characteristics

This report contains several derived statistical measures reflecting labor force and employment rates as of April 1993:

Labor force participation rate. The labor force is defined as those employed (E) plus those unemployed (U—i.e., those not-employed persons actively seeking work). The labor force participation rate (R_{LF}) is the ratio of the labor force to the population (P).

$$R_{LF} = (E+U) / P$$

Unemployment rate. The unemployment rate (R_U) is the ratio of those who are unemployed but seeking employment (U) to the total labor force (E+U).

$$R_U = U / (E+U)$$

S&E involuntarily out-of-field rate. The S&E involuntarily out-of-field rate (R_{IOF}) is the ratio of those who are working part-time but seeking full-time jobs (E_{PTS}), or who are working outside their degree field when an S&E job would be preferred (E_{NSP}), to total employment (E_T).

$$R_{IOF}=(E_{PTS} +E_{NSP})/E_T$$