Appendix E. Data Management and Data Issues

Here we describe the procedures we used to create the databases for the Phase I and II analyses presented in this report. We also discuss several issues involving the data. As noted in Section 2 of this report, both phases of this study used data from the Survey of Doctorate Recipients (SDR). The SDR is a nationally representative sample survey of doctorates earned in the United States in science and engineering (S&E).[1] The survey is conducted every two years and provides information on individual doctorate recipients' academic fields, career outcomes, and numerous personal characteristics (e.g., birth date, sex, race/ethnicity).[2] The SDR is longitudinal in the sense that individuals are tracked over time in successive survey waves throughout their careers as long as they remain in the sample frame.

Procedures for Creating Databases

We used a four-step process to create the databases for the Phase I and II analyses: (1) review documentation for outcome and control variables; (2) extract raw data from SDR files; (3) create variables for analyses; and (4) merge files across SDR waves.

Step 1 is relatively straightforward. We reviewed documentation for the SDR files to select the outcome variables of interest and the control variables used in the multivariate analyses. We also recorded the file positions for each selected variable.

During step 2 we extracted the raw data selected in step 1. Records were retained for individuals who reported full-time employment in academia and who earned their doctorates in S&E fields. Because the codes for some of the outcome and control variables are not consistently defined across SDR waves, we created separate files of the raw data for each SDR wave. After the data files were created, we computed summary statistics—including means and sample minima and maxima—for each of the raw variables. We then reviewed the summary statistics for potential errors in the raw data.

In step 3 we first wrote computer code for creating variables suitable for the analyses from the raw data. We then generated separate files of created variables for each SDR wave.[3] Finally, we computed summary statistics for each file and reviewed them for potential errors.

Finally, in step 4, we merged the files created in step 3 to create two large files of created variables—one each for Phases I and II. Merging the SDR files for the Phase I database was relatively straightforward because there was no requirement to link individual records across SDR waves. However, because the Phase II analyses are longitudinal in nature, we had to match records across files using doctorate recipients' identification numbers.[4] It was also necessary to create some of the Phase II analytical variables after the files were merged. These include the outcome, outcome status, and employment variables that require linked historical records for construction. As a final quality control check, we generated summary statistics for each of the two merged files and reviewed them for potential errors.

Data Issues

Issues involving the data, such as missing data, data errors, changes in the SDR survey instrument, constructing Phase II historical records, and limited control variables, surfaced during the course of this study.

Missing Data

Missing data occur in the samples we used for this study for two reasons. First, some individuals failed to complete the entire survey questionnaire. Second, the design of the SDR survey instrument has changed over time. As a result, some of the variables we used are available for some SDR waves but not for others.

[1] Although some SDR waves also include doctorates earned in business and the humanities, this study is limited to doctorates in S&E.

[2] A license is required to obtain SDR data in order to protect the anonymity of survey respondents and the confidentiality of their responses.

[3] Because the codes for some raw variables in the SDR files are not consistently defined across all waves, it was necessary to edit the computer code for the created variables accordingly.

[4] The Phase II data include individuals reporting full-time academic employment in the 1997 SDR wave, but the analyses require that historical records be constructed from previous SDR waves. See Section 2 of this report.

Our analyses excluded observations for which the outcome (i.e., the dependent) variables were missing; however, we did not discard observations for which control variables were missing (see Section 2). Instead, we adopted the approach of including "missing" dummy (i.e., dichotomous) variables as additional controls. This approach, which allowed us to retain larger samples, treats missing cases as special categories and allowed us to control for the marginal relations between missing characteristics and outcomes.

Data Errors

The data appear to be relatively free of errors. Our quality control measures detected only one apparent error in the raw data. We computed age at the time of the doctorate as the year of the doctorate minus the year of birth. This procedure yielded an illogical age for one individual, apparently because of an error in the birth date. We recorded age at the time of the doctorate as missing for this individual.

Changes in the SDR Survey Instrument

Changes in the SDR survey questionnaire have occurred over time. For example, the 1995 and 1997 SDR questionnaires ask for very detailed information on the number and ages of dependents in the family. Information on dependents reported in earlier waves, however, is less detailed. As a result, we were forced to limit our construction of family characteristics to two variables describing dependents, the number of children younger than age 6, and the number of children between the ages of 6 and 18. This was the most detailed common denominator that could be constructed for the family variables over the 1981 through 1997 SDR waves. Moreover, information on the characteristics of dependents is so sparse in SDR waves before 1981 that we excluded these from our analyses.

Constructing Phase II Historical Records

The Phase II analysis required us to identify the date at which key outcomes occurred in doctorate recipients' careers (i.e., dates of tenure and promotions to higher academic ranks).[5] Unfortunately, the SDR files simply indicate whether an individual is tenured (or has achieved a given academic rank) as of the date of the survey but do not indicate the date of tenure (or date of promotion). As a result, we had to search through all SDR records to determine first occurrences of reported tenure or employment in senior ranks. This procedure introduces the possibility of measurement error if data are missing at the dates of tenure or promotion (see Section 2).[6]

Limited Control Variables

The controls we have used in this study are limited by the data available in the SDR files. In our view, the most serious limitation is the lack of measures of productivity. We acknowledge that measures of teaching productivity and service to the institution and community are difficult to construct. However, we believe that measures of scholarly productivity—counts of articles and books published and papers presented at professional conferences—would have been useful. Apart from simple cumulative counts of publications, the timing of scholarly productivity is likely to be important. For example, establishing a scholarly record early in the postdoctoral career is likely to be an important criterion for earning tenure and for promotion to the associate professor rank at most academic institutions.

The 1995 SDR file does report measures of scholarship—the number articles published and papers presented.[7] The sample size from this single wave, however, is not adequate to estimate the models we have specified in this study.

[5] Dates for outcomes are required to construct variables measuring the time elapsed between earning the doctorate and tenure and promotions.

[6] Apart from the issue of potential measurement error, the design of the SDR files is somewhat awkward for use in longitudinal studies of career outcomes. The files indicate the states of outcomes as of the date of each survey wave but not when the outcomes first occurred. For example, the data for each survey wave indicate whether doctorate recipients are tenured at a given point in time but not when tenure first occurred. The same is true for promotions to higher academic ranks. Constructing the longitudinal variables also required matching individuals across files for different survey waves by respondent identification numbers.

[7] The 1995 wave also reports information on patent activity.