CHARACTERISTICS OF THE TOP R&D PERFORMING FIRMS IN THE U.S.:

EVIDENCE FROM THE SURVEY OF INDUSTRIAL R&D

by

Lucia Foster *
U.S. Bureau of the Census

and

Cheryl Grim *
U.S. Bureau of the Census

CES 10-33 September, 2010

The research program of the Center for Economic Studies (CES) produces a wide range of economic analyses to improve the statistical programs of the U.S. Census Bureau. Many of these analyses take the form of CES research papers. The papers have not undergone the review accorded Census Bureau publications and no endorsement should be inferred. Any opinions and conclusions expressed herein are those of the author(s) and do not necessarily represent the views of the U.S. Census Bureau. All results have been reviewed to ensure that no confidential information is disclosed. Republication in whole or part must be cleared with the authors.

To obtain information about the series, see www.ces.census.gov or contact Cheryl Grim, Editor, Discussion Papers, U.S. Census Bureau, Center for Economic Studies 2K130B, 4600 Silver Hill Road, Washington, DC 20233, CES.Papers.List@census.gov.

Abstract

Innovation drives economic growth and productivity growth, and as such, indicators of innovative activity such as research and development (R&D) expenditures are of paramount importance. We combine Census confidential microdata from two sources in order to examine the characteristics of the top R&D performing firms in the U.S. economy. We use the Survey of Industrial Research and Development (SIRD) to identify the top 200 R&D performing firms in 2003 and, to the extent possible, to trace the evolution of these firms from 1957 to 2007. The Longitudinal Business Database (LBD) further extends our knowledge about these firms and enables us to make comparisons to the U.S. economy. By linking the SIRD and the LBD we are able to create a detailed portrait of the evolution of the top R&D performing firms in the U.S.

* We thank seminar participants at the Center for Economic Studies for helpful comments. The authors are grateful to Paul Bailey for creating the initial datasets. Any opinions and conclusions expressed herein are those of the authors and do not necessarily represent the views of the U.S. Census Bureau. All results have been reviewed to ensure that no confidential information is disclosed.

1. Introduction

Innovation drives economic growth and productivity growth, and as such, developing reliable indicators of innovative activity is of paramount importance. Many national statistical agencies and independent researchers have attempted to develop indicators of innovative activity. While some surveys attempt to directly measure innovative activity (such as the European Union's Community Innovation Survey), other surveys attempt to measure inputs associated with innovative activity. Most of the research in the U.S. to date concerning innovation has focused on attempts to measure inputs.[1] Some of these efforts focus on counts of patents, for example, while others focus on capturing research and development activities. As an example of the latter, from 1953-2007, the National Science Foundation in conjunction with two U.S. statistical agencies collected data on research and development activities in the U.S. We use the data from this survey, the Survey of Industrial Research and Development (SIRD), to provide a detailed picture of the top R&D performing firms in the U.S. economy.

We are able to provide an unusually rich portrait of these top R&D performing firms by combining Census confidential microdata from two sources. The SIRD provides detailed information on R&D expenditures for a sample of firms, while the Longitudinal Business Database (LBD) provides a limited set of information for all firms. Moreover, the LBD enables us to supplement the firm-level information from the SIRD with establishment-level information. Linking the SIRD and the LBD allows us to create a detailed portrait of the evolution of the top R&D performing firms in the U.S.

We start by providing a detailed portrait of the top 200 U.S. R&D performing firms in 2003. We choose the level of total R&D expenditures as our metric for "top" R&D performers. An alternate measure of "top" R&D performers is the intensity of R&D effort (e.g., ratio of R&D expenditures to total sales revenue). However, we focus simply on the level of R&D expenditures as our measure since this provides a more direct link to the aggregate level of R&D expenditures in the U.S. There are some firms that have very high R&D intensity measures because they have very little output but do not contribute significantly to overall R&D activity in the U.S.

We measure the firm size, firm age, and firm structure of these top R&D performing firms. We compare these characteristics for the top performing firms to both other R&D performing firms and to all firms in the economy. To preview our findings, we find that the top R&D performing firms as compared to other R&D performing firms (or firms in general) are more likely to be large, older, multi-unit firms in the manufacturing sector.

We then look at the dynamics of these top 200 firms paying attention to their persistence and evolution. We find that R&D performing firms, and in particular, the top 200 R&D performing firms in 2003 are persistent relative to firms in the economy as a whole. Over half of the top 200 R&D performing firms in 2003 existed in the 1976 LBD, while only 20% of multi-unit firms in the whole economy in 2003 existed in the 1976 LBD.[2] R&D performers are also persistent relative to the whole economy; 30% of 2003

[1] Atrostic (2008) provides a detailed overview of attempts to measure innovative activity in the U.S.
[2] The LBD begins in 1976.

R&D performers and 14% of firms in the whole economy appear in the 1976 LBD. We consider three evolutionary paths for the top 200 R&D performers: those that were already large in terms of R&D activity at the start of our sample, those that were small but grew, and those that started during the sample and then grew. Of the top 200 R&D firms in 2003 that were also in existence in 1976, unfortunately not all have data on R&D expenditures for 1976. Of those with both 1976 and 2003 R&D expenditures data, we find that most were also top R&D performers in 1976 and that they managed to stay on top by more than doubling their real R&D expenditures. For those top 200 R&D firms in 2003, who existed in 1976 but were not in the top group, they made their way into the top 200 by 2003 not surprisingly through even more dramatic increases in R&D expenditures. For the smaller set of R&D top performers in 2003 who did not exist in 1976, we found that most of them started out as relatively small single units. These entrants then experienced rapid growth.

The paper is organized as follows. In the next section we provide a literature review summarizing results of existing work on the characteristics and evolution of R&D performing firms. Section 3 describes the data and measurement issues. Basic firm characteristics, including structure, size, age, industry, and location are described in Section 4. The main results regarding the dynamics of the top 200 R&D performing firms are presented in Section 5. Section 6 provides conclusions and some potential future extensions to our research.

2. Literature Review

While there are numerous strands in the economic literature concerning the characteristics of firms that successfully undertake R&D activity, we focus on those areas of the literature that we can address with our data. Recall that we not only want to provide a snapshot of the top R&D performing firms in 2003, we are also interested in their evolution over time. Thus we need to be able to measure these characteristics consistently over time. Specifically, the strengths in our data allow us to accurately measure over time: R&D expenditures, firm size (and structure), firm age, and industry.[3] We focus on two related strands in the literature on R&D activity: the impacts of firm size and market structure on R&D activities. We also note that the debate on firm size in other arenas has been transformed into a debate about firm age. The data are not as well suited to measure the exact location of R&D activities or of the technology of the firm. Therefore we do not attempt to measure these characteristics, nor do we include a summary of the vast literature in these areas of research.

2.1 Firm Size

The debate concerning the relationship between firm size and innovation started with Schumpeter's (1942) assertion that large firms are the engines of innovative activities. Arguments that large firms are more likely to engage in innovative activities often cite their relative advantage in terms of risk diversification and access to internal funds. Furthermore, when a large firm has significant market control, then the firm is

[3] The industry classification process used in the SIRD up until 2003 has some issues that make it less informative than the process that was started in 2004. This is discussed later in the paper.

more likely to be able to capture the output of the innovation. Many authors have offered counter arguments for why large firms are in fact less likely to be innovators. These arguments include that large firms face greater coordination issues and lack the flexibility needed to stay on the cutting edge of innovation. The discussion is similar to the debate concerning firm size and other economic activities such as job creation and destruction (see Jarmin, Haltiwanger, and Miranda (2010)). Some researchers question even the relevancy of this argument as the concept of a firm has less meaning when alliances across firms are used to engage in innovative activities. Teece (1992, p. 447) notes that "Discussions of the link between firm size and innovation are outmoded because the boundaries of the firm have become fuzzy in recent decades. Strategic alliances – constellations of bilateral agreements among firms – are increasingly necessary to support innovative activities."

Griliches (1980) examines the relationship between R&D expenditures and firm size using firm-level SIRD data from 1957-65. The sample in his paper is restricted to large R&D performing manufacturing companies (companies with at least 1,000 employees). The SIRD data were linked to Census of Manufactures data and Enterprise Statistics data ultimately yielding a sample of 883 companies. Company size is measured as the log of the average of the value added output for the company in 1957 and 1963. R&D expenditures are measured as the log of the R&D total expenditures cumulated over 1957-1965. The baseline regression results suggest that there is a positive relationship between R&D expenditures and size.

However, once Griliches includes controls for data issues (imputations) and differences in specialization, this relationship disappears. Similarly, Cohen *et al* (1987) find that the positive relationship between firm size and R&D expenditures disappears once they control for industry (and remove outliers). They use line of business data for about 300 large companies in the U.S. and while they find that firm size does not have a significant impact of the level of R&D intensity, they find that business unit size does have a positive impact on the probability that the firm is an R&D performer. Akcigit and Kerr (2010) using both the SIRD and the LBD find that R&D expenditures rise with firm size, but that R&D intensity (R&D expenditures as a share of sales or total employment) declines with firm size.

A series of papers examines the relationship between firm size and R&D more closely allowing for differences in types of R&D. Mansfield (1981) notes that one of the most important characteristics of R&D expenditures is their heterogeneity and therefore it is important to take this into account when considering the relationship of R&D to firm size. He finds that large firms carry out disproportionately more basic R&D but disproportionately less R&D aimed at new products and processes. Cohen and Klepper (1996) refine the argument concerning firm size and R&D activity to focus on the relationship between business unit size and the composition of R&D activity in terms of product versus process innovation. The important difference between product and process innovation relates to the appropriability of the innovation. It is assumed that process innovations cannot be sold in disembodied form to other firms. Thus all benefits must be captured within the firm. Firms that are larger can spread these costs over a larger amount of output. They find that large firms have a cost-advantage when it comes to R&D

relative to small firms and that this advantage is especially evident for process innovations.

Akcigit and Kerr (2010) also categorize R&D into two components that are related to those of Klepper and Cohen but differ slightly. Exploration R&D represents a firm's attempt to create a new product or become the leader in a market. Exploitation R&D represents a firm's refinements to improve existing product lines. Akcigit and Kerr link the SIRD to patent data which enables them to distinguish between these two types of R&D. Exploration R&D is measured by patents that do not self-cite; while the majority of citations in exploitation R&D patents are self-cites. They find that exploitation R&D is associated with larger scale firms, while exploration R&D does not have as strong a firm-size relationship.

The debate on business dynamics and firm size has shifted to firm age. Jarmin, Haltiwanger, and Miranda (2010) use the LBD to reexamine the relationship between firm size and employment growth. While the existing literature had found a positive relationship between firm size and employment growth, they find that this relationship disappears once they control for firm age. They find that most new jobs are created by *young* businesses and that young businesses also have a high job destruction rate. Given these findings, we consider firm age along with firm size in this paper. One model prediction in Akcigit and Kerr (2010) is that younger firms are more innovative than older firms. They posit that this is because "[b]y their nature, the smaller firms and new entrants are focused on exploration innovations (p.2)." They find evidence consistent with this in their data, but do not formally look at the impact of age noting that the relationship in their model is driven by the positive relationship between firm size and firm age.

2.2 Market Structure

There is also considerable debate concerning the role of market structure on a firm's innovations. As with the debate concerning firm size, some of this debate can be traced back to Schumpeter. Arguments that higher concentration markets lead to more R&D spending and innovative activities include a reduction in uncertainty, more accessibility to capital markets, and easier appropriability of returns from R&D investment. On the other side, arguments that higher concentration leads to less R&D spending and innovative activities stress that competition encourages risk-taking and innovation. Acs and Audretsch (1987, 1988) examine the relationship between innovations and a number of firm characteristics. Innovations are measured using the Small Business Administration's database of innovations that is derived from information listed in over 100 technology, engineering, and trade journals. Thus, they are measuring the outcomes of innovative activities rather than one of its inputs, R&D expenditures.[4] They find that lower levels of concentration (as measured using four-firm concentration ratios) are associated with higher levels of innovation activity. Moreover, they find that there is an interaction between firm size and market structure when analyzing innovations. Large firms have an innovation advantage in relatively concentrated

[4] The correlation between their measure of innovations and total R&D expenditures is only 0.48 but when focusing on company R&D expenditures is 0.75. We rely on total R&D expenditures in this paper.

markets, while small firms have an innovation advantage when markets are more competitive.

Levin *et al* (1985) find that the relevant factor is not market concentration but is instead the more fundamental characteristics of the underlying technology. Specifically, they are interested in technological opportunity (which combines concepts such as closeness to science, technology within the industry, and the maturity of the industry) and technological appropriability (the ability to capture the returns to an innovation). Using line of business level data collected by Levin, they find that the effect of market concentration on R&D intensity essentially disappears when one controls for technological characteristics.

It is beyond the scope of this paper to examine whether these R&D expenditures are translated into innovations and ultimately into higher economic growth and productivity growth. Clearly not all R&D expenditures results in innovations and not all innovations result in productivity growth. Akcigit and Kerr (2010) use patent citation data to designate R&D performers as having produced an innovation. They classify these innovations in terms of their quality based upon whether the patent citations are mostly self-referential or are outward-focused. Combining this data with the SIRD, they find a positive relationship between firm employment growth and patents and that firm employment growth is higher when more of a firm's innovations are higher quality. Griliches (1980, 1986) used the SIRD and found that R&D expenditures contribute positively to productivity growth. Moreover, Griliches found that basic research and company-based expenditures on R&D (as opposed to applied research and federal expenditures) were especially strong as productivity determinants. For now, we sidestep the question about whether these R&D expenditures result in positive outcomes. Future research using our micro-level data will allow us to see whether R&D expenditures lead to economic and productivity growth.

3. Data and Measurement Issues

The empirical exercises in this paper use three micro-level confidential datasets. The first dataset contains the micro-level data collected by the Census Bureau from 1972 to 2007 as part of the SIRD. We supplement this R&D data with earlier SIRD data from the Griliches historical dataset. The third dataset is the LBD which was developed by economists at the Census Bureau. In this section of the paper, we describe the relevant features of these datasets. As with all Census confidential data, results reported from empirical exercises using these data must pass strict disclosure avoidance review. Operationally, this means that no firms will be mentioned by name and all identifying information will be suppressed in this paper.

3.1 Survey of Industrial Research and Development

The SIRD collects information on R&D as performed by businesses (as opposed to academia or government) in the U.S. As such, it is estimated that the survey covers more than two-thirds of funds spent on R&D in the U.S.[5] The SIRD collects data on total

[5] National Research Council (2005), page 48.

R&D expenditures and details of these expenditures by type (basic research, applied research, and development) and source of funding (Federal R&D funds versus company R&D funds). It also collects data on items such as net sales, total domestic employment, distribution of R&D by state, number of scientists and engineers, wages and salaries of R&D personnel, costs of materials and supplies consumed, and depreciation on R&D property and equipment.

The SIRD was collected as a joint partnership between the sponsoring agency, the National Science Foundation (NSF), and the Census Bureau beginning in 1957 and ending in 2007. Starting with the reference year 2008, the SIRD survey was substantially redesigned and renamed the Business Research and Development and Innovation Survey (BRDIS). The Center for Economic Studies currently has access to the full set of SIRD microdata starting in 1972 and going through 2007. The SIRD microdata have been used in research examining the relationship between R&D expenditures and firm productivity; buyouts, mergers and acquisitions; patenting activity; learning by doing; and import competition.[6]

We start by describing in detail how the SIRD sample is derived. We then describe the SIRD data, paying attention to issues that have been highlighted by others. Issues concerning comparability over time in the way that the sample was drawn and other issues concerning data drive many of our decisions in this paper.

3.1.1. Sample Selection for the SIRD

We summarize sample selection and form distribution for the reference year 2003 in Table 1. In the table, the left column shows the activity while the right column shows the number of companies. The frame is the Census Bureau's business register. For companies with more than one establishment, the data are summed to the company level and the resulting company record is used to select the sample and to process and tabulate the survey data. The frame is partitioned into three groups: (1) companies known to conduct R&D in the previous 5 years; (2) companies that reported zero R&D in the previous 5 years; and (3) companies for which R&D activity was uncertain. As is clear from the table, the third group dwarfs the first two groups.[7]

Groups 1 and 3 are treated similarly for sample selection. They both have two treatments, the only difference between the two groups is the measure of size: for group 1 it is R&D expenditures while for group 3 it is payroll. The first treatment for these two groups is certainty companies. For group 1, these are companies with prior R&D of $3 million or more. For group 3, these are companies in top 50 of their state by payroll. The second treatment for groups 1 and 3, is probability related to size. Group 2 is subject to simple random sampling.

---

[6] Productivity: Griliches (1980, 1986), Siegel and Lichtenberg (1989), Nguyen and Kokkelenberg (1991), Adams and Jaffe (1994). Buyouts, mergers and acquisitions: Lichtenberg and Siegel (1990a, 1990b), Long and Ravenscraft (1993a, 1993b). Patenting activity: Kerr and Fu (2006), Balasubramanian and Sivadasan (forthcoming). Learning by doing: Jarmin (1996). Import competition: Scherer and Huh (1992).

[7] Note that the total number of companies in these three groups does not sum to the total number of companies on the business register for a couple of reasons. The most important reason is that the SIRD only samples firms with five or more employees. Other exclusions from the SIRD include agricultural firms and non-profit firms.

The table also shows the rules for the distribution of survey forms. The SIRD is collected using two forms: a long form (RD-1) and a short form (RD-1A). The long form is sent to companies with more than $3 million in R&D. Notice that only about 30% of certainty companies get the long form. The short form is sent to all others. Generally speaking, there has been about a one in ten chance of finding an R&D performing firm in the non-certainty portion of the sample.

3.1.2. SIRD Data Issues

In this section, we focus on SIRD data issues that impact our research. These issues have been identified by previous users of the SIRD microdata and by the Committee on National Statistics (CNSTAT).[8] We categorize these issues into five broad groups: (1) sample scope, (2) coverage, (3) industry coding, (4) location of R&D expenditures, and (5) mergers and acquisitions. As a result of these issues, the SIRD was completely redesigned and is, as noted above, now known as the Business R&D and Innovation Survey (BRDIS).

The first issue concerns the consistency of the scope and coverage over time. The frame of the SIRD is the Census Bureau's Business Register. Over time, the sample drawn from this frame has expanded. The SIRD originally covered only the manufacturing sector, but now includes all for-profit R&D performing firms that operate in the U.S.[9] Sampling rules have also changed over time. For example, in the early years, the SIRD used a panel drawn every five years that was not refreshed. Thus, it was impossible for the SIRD to capture new firms within a panel. In 1992, the SIRD switched to using an annual panel and the sample size was approximately doubled.[10] In addition, due to issues related to the collection of data on R&D by state (see below), a certainty component was added so that the 2003 SIRD includes with certainty the largest 50 firms (by payroll) for each of the 50 states. Therefore, the number of firms in the survey varies over time for reasons unrelated to the size of the underlying population. One of the challenges in using the SIRD is that these changes in the survey make it difficult to make meaningful comparisons of aggregate statistics over time. Our focus on the top 200 firms partly mitigates this problem as does our reliance on the LBD to provide additional aggregate statistics.

The second issue concerns coverage. The SIRD is collected via a long form and a short form. In recent years, the two forms are both used in a given year: the short-form for non-certainty cases and the long-form for certainty cases. However, in earlier years the short-form was used for certainty cases in even years and the long form for certainty cases in odd years. Thus some data are collected only for long form cases.[11] Additionally,

---

[8] There have been about 20 projects that have used the SIRD microdata via the Census Bureau's Research Data Centers. CNSTAT convened a panel in 2003 to conduct an in-depth review of the R&D statistics program in the SRS Division of NSF. The resulting review was published as a book, *Measuring Research and Development Expenditures* (National Research Council, 2005).

[9] Since 1976, the frame has been the Census Bureau's Business Register, previously known as the Standard Statistical Establishment List (SSEL). See National Research Council (2005), pp. 63 and 86.

[10] "The 1992 redesign effort updated the operations of the survey to reflect many aspects of the changed environment and correct some of the most severe deficiencies, which had rendered the survey results quite misleading over time." National Research Council (2005), page 49.

[11] In 2003, the long form had 15 questions, while the short form had 10 questions.

not all questions are asked every year. Only a core set of data are collected every year: domestic net sales and receipts, domestic employment, number of R&D scientists and engineers, total costs incurred for R&D, total costs incurred for R&D from Federal funds, and the cost of R&D by state. These are also the only questions that were deemed mandatory for most of the survey years. While of interest, the non-mandatory questions are often sparsely populated.[12] In this paper, we focus on the questions that were mandatory over all years. We use all of the available microdata for these questions and thus use reported, edited, and imputed data.

The third issue concerns the industry coding in the SIRD. The SIRD does not collect information on industry; instead an industry code for a firm is derived from establishment-level data from the Business Register using a hierarchical algorithm. This algorithm assigns a four-digit NAICS code based upon establishment-level payroll.[13] Thus all of the R&D expenditures in a firm, even a very diverse firm, are assigned to only one industry. In addition to suppressing the diversity of R&D activity, this process can also result in striking changes in R&D activity by sector when large firms are reassigned.[14] Furthermore, as traditional manufacturing firms increasingly include wholesaling activity, this method of allocating R&D activity based upon payroll resulted in the allocation of an increasing amount of R&D activity to the wholesale trade sector. During the processing of the 2004 SIRD, the increasing allocation of R&D activity to this sector came under scrutiny. As a result of this scrutiny, much of the R&D for wholesale and management sectors was redistributed.[15] We do not apply this type of redistribution to the 2003 microdata for our analyses for disclosure avoidance reasons. However, in future work we plan to use these improved industry codes when creating a measure of industry concentration since we are pooling industries in that case.

The fourth issue concerns the SIRD data on the location of R&D expenditures by state. Unlike the industry codes, the SIRD does collect information on location; however, it is only collected on the long form. Moreover, prior to 2001, it was not a mandatory data item. Some observers have noted that state R&D expenditure totals are likely subject to issues related to the small underlying population of R&D performing firms in the states. These issues mean there are large variations in state totals over time that reflect the impact of small populations being given large weights. For this reason, the 2003 SIRD includes with certainty the largest 50 firms (by payroll) for each of the 50 states.[16]

---

[12] The non-mandatory data include whether the R&D is basic or applied, other costs at the firm including the wages and salaries of R&D personnel, costs of materials and supplies consumed, and depreciation on R&D property and equipment.

[13] Specifically, the firm is first assigned a two-digit NAICS code based upon the sector (two-digit NAICS) in which it has the most payroll as aggregated from its component establishments. The firm is then assigned a three-digit NAICS code based upon the subsector (three-digit NAICS) that accounted for the highest percentage of payroll in the dominant two-digit sector. This process is then repeated at the four-digit NAICS level.

[14] National Research Council (2005), page 31.

[15] At the same time, industry classifications for the information, professional, scientific, and technical services, and management of firms and enterprises sectors were also examined. Firms in all of these sectors were subject to a manual review in which firms were reclassified based on primary R&D activity usually as defined by their primary product or service. For more details, see National Science Foundation (2009b), Appendix A.

[16] National Research Council (2005), pp. 63 and 181.

The final issue concerns the impact of mergers and acquisitions on the SIRD data. Because the SIRD is a company-level as opposed to an establishment-level survey, mergers and acquisitions can cause spurious changes in R&D expenditures over time. Scherer and Huh (1992) attempt to address this issue by creating a smoothed series. Griliches and Mairesse (1981), when using Compustat data, try to mitigate the impact of mergers and acquisitions on the data through a variety of methods (e.g., dropping the company, allowing the company to exist before and after mergers and acquisitions). They find that about 20% of their firms have jumps in the number of plants, employees, and sales suggesting that the company has undergone mergers and acquisitions. We have not yet addressed this issue, but we plan to do so in the future.

3.1.3. Properties of the Sample from the SIRD

These issues then lead us to our basic strategy of focusing on the core activities of large R&D performers. That is, we focus mostly on data from mandatory questions and for companies that are likely to have been certainty companies. As the CNSTAT book notes: "Today's R&D survey is, in reality, two surveys in one. One survey which has existed from the earliest days of the data collection is a survey of companies thought to have the largest R&D expenditures." We have essentially chosen to focus on a subset of this survey of large R&D companies. This is not a novel strategy when using the SIRD. As we are about to discuss in the next section, one of the first research efforts using the SIRD, the Griliches papers, focused on large R&D performers. Numerous other research papers using the microdata from the SIRD have also focused on large companies.

The main variable that we use in this paper is total R&D expenditures. Total R&D expenditures include the following: basic research (the planned, systematic pursuit of new knowledge or understanding toward general application), applied research (the acquisition of knowledge or understanding to meet a specific, recognized need), and development (the application of knowledge or understanding toward the production or improvement of a product, service, process, or method). This variable does not include quality control, routine product testing, market research, sales promotion, routine technical services, or research in the social sciences. Total R&D expenditures are for R&D that is performed both within the respondents' company and within the U.S. (see the Appendix for more details).

Figure 1 shows the level of total R&D expenditures in constant dollars as measured by the SIRD from 1953 to 2007 for the published data (solid line) and the underlying available microdata (dashed line).[17] As is evident from the plot, R&D expenditures in the U.S. have been growing over time. The underlying microdata used to create these published statistics follow the same trend and come close to replicating these published totals in most years, but there are some years where they noticeably diverge. Generally speaking, differences between the published and microdata reflect revisions to the SIRD data. These revisions were applied to the published data but were not carried back to the original micro-level source data. Revisions to the data can occur for a variety of reasons. We note the following three examples as described in NSF (2007b). For the 2003 data, a very large company did not respond and so had its data imputed. Upon

[17] Throughout this paper, we convert current dollars to constant dollars using the Gross Domestic Product implicit price deflator employed by the National Science Foundation in their SIRD publications.

comparing this to publicly available data, it was realized that the imputation was too low. For the 1999 data, a large company was given a large weight. The next year it was picked as a certainty company and its weight dropped to one, causing a large spurious change in total R&D expenditures for that company. A revision in methodology caused a "dramatic" revision to the 1991 data. (This was the change in scope discussed earlier where the SIRD started to use an annually refreshed panel and the sample size was doubled). We cannot incorporate these revisions into the micro-level data, so we are mindful of suspiciously large changes in our data.

The SIRD microdata contain a firm identifier from the Business Register. Thus it is possible to link SIRD microdata over time and to microdata from other Census datasets. The firms in the SIRD were linked over time using consistent firm identifiers that were derived from Census firm identifiers but modified to account for changes in Census processing. The files were also cleaned to take in to account differences in retention policies over the years (that is, in some years, prior year observations were stored with the current year).[18] There are close to 500,000 firm-year observations representing over 300,000 firms in the longitudinal SIRD dataset spanning 1972-2007. In 2003, the SIRD contains approximately 31,000 firms with approximately $200 billion in R&D expenditures.[19]

The firms in the SIRD include a number of firms that do not report positive R&D expenditures. In this paper we distinguish between all of the firms in the SIRD ("R&D sample") and firms in the SIRD who report positive total R&D funds ("R&D performers"). Only about 20% of the firms in the SIRD reported positive total R&D expenditures in 2003. One of the difficult issues faced is identifying firms who perform a significant amount of R&D. In order to be included in the certainty sample in 2003, a firm had to have R&D expenditures of at least $3 million in the 2002 SIRD or to be in top 50 of their strata or state in terms of payroll.[20] The SIRD survey "imposes a burden on nine firms to find just one that does reportable R&D."[21] Consistent with this issue of finding a needle in a haystack, close to 85% of the firms in the longitudinal file appear for only one year in the SIRD. It is important to keep in mind that firms may exit the sample for many reasons other than the death of the firm.

The SIRD microdata concerning non-activity is less reliable in earlier years due to the way missing and zero values were captured. Thus, we restrict our analysis of non-activity to 1994 and later. We find that only about 20% of the firms in our sample actually reported positive R&D expenditures for 1994-2007. Focusing on a slightly more limited set of years, over 60% of firms in 1994-2006 appear in only one year.[22] Of these "one and done" firms, only 3% have positive R&D expenditures. In order to address the issue of locating firms with positive R&D expenditures, the Census Bureau recently

[18] Future users of the data can see technical details concerning data for this paper in Foster and Grim (forthcoming).
[19] See page 182 and Table 1 of National Science Foundation (2007b).
[20] See NSF (2007b), p. 181. The cut-off for R&D expenditures for certainty cases has changed over time. See National Research Council (2005) for a detailed timeline of changes in the SIRD.
[21] National Research Council (2005), page 63.
[22] We do not include 2007, our last year of data, in this analysis of "one and done" firms since firms that appear only in 2007 may appear in later years.

added a question to its annual Report of Organization Survey (ROS) concerning R&D expenditures that has provided useful information in building the survey frame for the SIRD.[23]

While there are a significant number of "one and done" firms, we also see many firms that appear in the SIRD across multiple years. About 10% of firms in the SIRD appear for between 2 and 6 years.[24] At the other end of the distribution, close to 500 firms appear in every year of the SIRD.

3.2 Griliches Dataset of Partial Historical SIRD 1957-1977

SIRD company-level data back to 1957 was recently recovered for a subset of firms. In the mid-1960s, Zvi Griliches was approached by the Census Bureau and the National Science Foundation to work on a project analyzing historical data on industrial R&D. Griliches led efforts to develop a longitudinally-linked dataset containing 1957-1965 company-level data from the SIRD matched to data from the 1958 and 1963 Census of Manufactures (CM) and Enterprise Statistics (ES). Griliches (1980) details this effort and discusses several limitations of the data.

Griliches returned to the Census Bureau to extend his work in the early 1980s. Unfortunately, the data from his original project had been lost (Griliches, 1986). Griliches designed a new dataset over a longer time period. The new dataset contained 1957-1977 firm-level data from the SIRD matched to data from the 1962, 1967, and 1977 ES and data from the 1967 and 1972 CM. The sample was limited to certainty companies from the 1972 SIRD, and the final sample contains approximately 1,100 companies. The data underlying the findings in Griliches (1986) was recently recovered as part of a larger team effort to recover historical micro-level Census data.[25]

Fortunately, we are able to link the Griliches dataset to the main SIRD dataset using the firm identifiers. In this way, we are able to extend our dataset back to 1957 for a subset of firms (see section 3.4 for a detailed discussion).

3.3 Longitudinal Business Database

The Business Register (BR) contains nearly all non-farm private business establishments in the U.S. economy. In 2003, there were approximately 6.1 million establishments in the BR (Fairman et al., 2008). The primary source of information on businesses in the BR comes from the Business Master File and the Business Income Tax and Payroll Tax forms from the Internal Revenue Service. In addition, information from the Census Bureau's Economic Census and ROS are included on the BR. Finally, additional information concerning industry and location comes from the Bureau of Labor Statistics and the Social Security Administration. The BR includes information on employment, payroll, industry, and location. The LBD links the BR over time starting

---

[23] The R&D question was first added to the ROS in 2003. It has appeared on the ROS in every year since then except in 2006. R&D questions were also added to the 2007 Economic Census.

[24] Firms that appear in multiple years of the SIRD may have gaps in their data (e.g., appear in the data, disappear from the data, and then reappear in a later year). Gaps in the data are expected given the nature of sample selection for the SIRD. Of the firms that appear in multiple years in our data, 51% have no gaps in their data.

[25] See Becker and Grim (2009) for a discussion of the historical data recovery effort.

with the year 1976.[26]

Jarmin and Miranda (2002) paid particular attention to maintaining high quality links in the establishment data over time when they created the LBD. The longitudinal linkages are created using a combination of numeric establishment identifiers and name and address. Periods of inactivity for establishments are allowed in the LBD thus mitigating the problem of false births and deaths.

The SIRD is a survey so it is not possible to determine whether or not a firm that is missing from the SIRD in a given year was in existence in that year. Matching SIRD firms to the LBD enables us to determine the status of a firm in a given year by relying on the existence of its component establishments. Since the SIRD extends back to 1972, while the LBD only goes back to 1976, SIRD firm-year observations from 1972 through 1975 cannot be matched to the LBD. Moreover, the LBD contains information that allows us to create a measure of firm age (albeit one with left-censoring). Matching SIRD firms to the LBD also allows us to add data on the establishments within R&D performing firms to our analysis.

Given that the LBD represents the universe of operating establishments, we expect the match rate between the SIRD and the LBD to be high. We do not expect that match rate to be perfect for a given year since there may be instances when changes in a firm are captured more quickly in one source than in the other. There are two types of firms in Census data: firms with more than one establishment (multi-unit firms) and firms with only one establishment (single unit firms). The matching algorithm used to match the firms from the SIRD to the establishments on the LBD differs for single units and multi-units. For the years 1976 to 2007, we are able to match approximately 96% of the single-unit firm-year observations and 99% of the multi-unit firm-year observations from the SIRD to the LBD.

3.4 Identifying the Top 200 R&D Firms

Before focusing on the top 200 R&D performing firms for one particular reference year, it is helpful to look at the characteristics of the top 200 R&D performing firms for our main sample period (1972-2007). We identify the top 200 R&D performing firms based upon their total R&D expenditures in each year. Figure 2 shows the published level of total R&D expenditures in constant dollars as measured by the SIRD from 1953 to 2007 (solid line) and the R&D expenditures of top 200 firms in each year for comparison (dashed line). The top 200 firms accounted for close to the total amount of R&D expenditures until about 1985, since then the top 200 firms have accounted for a decreasing share of total R&D expenditures. Nevertheless, the top 200 firms in 2003 still account for about two-thirds of the $200 billion R&D expenditures by businesses.[27]

We now focus on the top 200 R&D performing firms in 2003. There are over 4,500 firm-year observations for the top 200 firms in the longitudinal SIRD dataset spanning 1972-2007. Relative to the sample as a whole, these top 200 firms appear in

[26] The LBD uses information from the BR starting in 1975. However, since the LBD requires prior year information on a firm, the LBD database files begin in 1976.
[27] "It is estimated that the largest 100 firms account for 54 percent of R&D spending." National Research Council (2005), page 20.

many more years of the SIRD. In fact, approximately 40% of the top 200 firms appear in every year and about 85% appear in at least 10 years of the longitudinal SIRD dataset. There are a few instances of gaps in the data, but for the most part, the firms appear in multiple contiguous years. For convenience, we refer to these top 200 firms as defined in 2003 in the rest of this paper as the "Top2003" firms (since they are the top 200 as defined at the fixed point in time of 2003).

Figure 3 shows the number of Top2003 firms in the SIRD from 1972 to 2007 (solid line) and the number of Top2003 firms that are in the SIRD *and* among the top 200 R&D performers in that year (dashed line).[28] Many Top2003 firms that were in the SIRD sample historically were also in the top 200 R&D performers in earlier years. Also, there is a small, but notable, drop off in the number of Top2003 firms from 2003 to 2007. This drop is primarily attributable to merger and acquisition activity.

Matching the Top2003 to the LBD yields a high match rate. In some cases, a firm that appears in one year in the SIRD does not appear in that same year in the LBD, but we are able to find it in other years of the LBD. This leads to gaps in the data from the LBD.[29] Thus, our establishment-year data from the LBD includes some missing values for our Top2003 firms.

We are able to match around 40% of the Top2003 firms to the historical Griliches dataset. Over 90% of the matched firms (over 35% of all Top2003 firms) had positive R&D expenditures in 1957. Since the recovered data do not contain the universe of firms or even all SIRD firms, we are limited in our ability to draw conclusions from our match of Top2003 firms to the historical data. However, we can say that over 35% of Top2003 firms have been in existence and doing R&D for over 45 years.

4. Snapshot of Top R&D Performing Firms

In this section, we provide a snapshot of the Top2003 firms in 2003. We start by providing some summary statistics on these firms and then provide details on their size, age, and structure. We compare these characteristics to all of the other R&D performing firms (firms in the SIRD reporting positive total R&D funds) in 2003.[30] We also compare these firms to all firms in the economy and all multi-unit firms in the economy using data from the LBD.[31]

Table 2 shows summary statistics from the SIRD for R&D performers and Top2003 firms in 2003. Mean R&D expenditures for Top2003 firms are much larger than for R&D performers in general, with Top2003 firms accounting for approximately 65%

[28] There are large steps in the solid line. There was a new panel drawn between 1980 and 1981. Between 1991 and 1992, a new methodology with a larger sample size and a new panel were implemented in the SIRD. We are investigating the jump from 1985 to 1986. According to the documentation we have, there was a new panel drawn between 1986 and 1987, so the jump occurs a year earlier than expected. See NSF (1995).

[29] We attempted to fill in these data gaps using various imputation measures but found that we introduced outliers by doing so.

[30] All statistics for R&D performers in this paper are weighted using SIRD sample weights.

[31] Whole economy data are limited to firms with at least five employees to be consistent with the SIRD sample frame.

of total R&D expenditures in 2003. Top2003 firms are also significantly larger than R&D performers in general in terms of sales and employment. Over half of the scientists and engineers in R&D performing firms work for Top2003 firms. Figure 4 shows total R&D expenditures for all firms (based on the published data) and for Top2003 firms. Top2003 numbers prior to 1972 are based on our match of Top2003 firms to the historical Griliches data. This figure also shows the number of Top2003 firms that are in the SIRD in each year. The share of total R&D expenditures accounted for by Top2003 firms has varied over time, but generally speaking this group of firms accounts for at least 40% of total R&D expenditures over our main sample period (1972-2007).

Manufacturing firms account for 60% of R&D expenditures in 2003 and about 67% of the R&D expenditures of the Top2003 firms. Within manufacturing, the industries with the largest R&D expenditures in 2003 are computer and electronic products, transportation equipment, and chemicals. These three industries also comprise the majority of R&D expenditures in Top2003 firms. Not surprisingly, given the known issues with coding wholesale trade, the next largest concentration of firms in Top2003 firms is in wholesale trade. Firms in the Professional, Scientific, and Technical Services (NAICS 54, hereafter PST Services) account for about 14% of total R&D expenditures. In Top2003 firms, PST Services firms account for 7% of total R&D expenditures. In order to be able to compare to the general economy, we use employment as the measure of economic activity.[32] In 2003, employment in manufacturing accounted for about 65% of employment of Top2003 firms, 57% of employment of R&D performers, but only 17% of employment in multi-units and 14% of employment in the economy. Interestingly, the percent of employment accounted for by employment in PST Services is essentially constant at about 5% across the Top2003, R&D performers, multi-units, and the whole economy.

4.1 Firm Size

In keeping with much of the empirical literature that uses Census micro-level data, we measure firm size by the number of employees at the firm (some work uses firm sales instead). [33] Using this measure of firm size, it is clear that the Top2003 firms are larger than their R&D performing counterparts. As shown in Table 2, the median Top2003 firm has over 11,000 employees, while the median R&D performing firm has 28 employees.[34] We also considered analyzing a measure of R&D intensity. Our preliminary work creating this variable uncovered some issues with outliers. Akcigit and Kerr (2010) seemed to have faced similar issues given that their measures of R&D intensity (whether using sales or employment) have their 1% tails trimmed.

When comparing firm size across groups it is important to keep in mind differences related to technology of the firm. One way to crudely measure technology is

[32] We create firm-level industry classifications for the LBD using the same methodology as is applied to the SIRD in 2003.

[33] For example, Griliches (1980) uses value-added output as a measure of firm size. Acs and Audretsch (1987) and Akcigit and Kerr (2010) create alternate measures of innovation rates that use sales in the denominator rather than employment. Akcigit and Kerr (2010) also categorize firms by the number of their establishments.

[34] To ensure no confidential information is disclosed, reported medians are "fuzzy" medians, where the "fuzzy" median is the mean of all observations between the 45th and 55th percentiles.

by industry. Other work has shown that R&D performers in the service sector are noticeably smaller than R&D performers in the manufacturing sector: "Some 43 percent of currently measured R&D in the service sector was performed by firms with fewer than 500 employees, compared with a contribution of about 8 percent by smaller firms in the manufacturing sector." (National Research Council (2005), p.24). Below we crudely control for these differences and in our regressions we always control for industry fixed effects.

Holding technology (industry) constant, we find that the Top2003 firms are very large and R&D performing firms are large relative to the general population. Here we focus on two of the most important sectors for R&D expenditures: Manufacturing and PST Services.[35] For the manufacturing sector in the 2003, median firm sizes are the following: over 16,000 employees for Top2003, 153 employees for all multi-units in the economy, 46 employees for R&D performers, and 15 employees for the whole economy. R&D performers in the PST Services sector are also large relative to the general population. The median firm size in this industry for a Top2003 firm is over 1,000 employees, a multi-unit firm is 43, an R&D performing firm is 22 employees, and a firm in the whole economy is 9 employees.

Recall that some researchers noted that it was important to take into account the type of R&D activity at firms when looking at the relationship between firm size and R&D activity. While we do not have similar measures of the type of R&D activity, the SIRD does collect data on three types of activities: basic R&D, applied R&D, and development. In future drafts of this paper, we will add some analyses of these types.

4.2 Firm Age

There are multiple ways to define a firm's age. Age could be defined from the first appearance of the firm identifier or based on the age of the oldest establishment associated with the firm. Following Jarmin, Haltiwanger, and Miranda (2010), we define firm age as the age of the oldest establishment associated with the firm in the first year the firm identifier appears in the LBD.[36] Since the LBD begins in 1976, the oldest possible age for a 2003 firm is 27 years. As one might expect, top performing R&D firms tend to be older firms. As shown in Table 3, in 2003, the median age of firms in the LBD is 11 years old and for the subset of multi-unit firms is 25 years old. The median age of all firms with positive R&D expenditures in 2003 is 16 and the median age of the Top2003 firms is censored at 27.

Figure 5 shows the percent of firms by age category for the whole economy (all firms in the LBD), for multi-units only (all multi-unit firms in the LBD), for firms with positive R&D in 2003, and the Top2003 firms for the year 2003. Top2003 firms are older firms. Close to 65% of Top2003 firms are 27 years old or older, while only about 6% of the Top2003 firms are 10 years old or younger. In the whole economy, only around 17%

[35] We do not make comparisons for firms in Wholesale Services because, as discussed in Section 3.1.2, those industry codes are problematic.

[36] Establishment age is defined as the first appearance of an establishment in the LBD with positive employment. We require positive employment because we focus on employment as a measure of size in this paper. Note that because of the way employment is defined in the LBD (employment as of March 12th), it is possible for an establishment to have zero employment and positive payroll in a given year.

of firms are 27 years old or older, while about 50% of firms in the whole economy are 10 years old or younger. A more apt comparison is between the ages of the Top2003 firms and the ages of multi-unit firms in the LBD. About 20% of multi-unit firms in the LBD are 10 years old or younger, while only about 6% of the Top2003 firms are 10 years old or younger. Top2003 firms are also older than firms with positive R&D expenditures in 2003; about 25% of these firms are 27 years old or older.

4.3 Firm Structure

The link to the LBD enables us to learn more about the structure of the Top2003 firms. Table 3 shows summary statistics from the SIRD-LBD matched data for 2003. While some of the Top2003 firms appear to be single unit firms, it is probably more correct to think of all of the Top2003 firms as multi-unit firms. The few firms that are technically single units appear to have complex relationships with existing firms that make them closer in spirit to multi-unit firms. Given the way that the SIRD sample is drawn, we would expect many more of the firms to be multi-units than is the case in the general population. Thus, it is not surprising that about one-third of the firms in the SIRD sample are multi-units, as compared to in the economy as a whole where multi-units comprise about 8% of all firms in 2003. Multi-unit firms account for 21% of R&D performing firms in 2003.

A more relevant statistic is the economic activity associated with multi-unit firms. We have noted that it is reasonable to say close to 100% of all R&D expenditures of the Top2003 firms are associated with multi-unit firms. More generally, approximately 87% of all R&D expenditures in 2003 are associated with multi-unit firms. When comparing to the economy as a whole, we use employment as our metric of economic activity. Employment at multi-unit firms in 2003 accounts for about 100% for Top2003 firms, 92% for R&D performers, and 57% for all firms. Thus, the economic activity of Top2003 firms is more concentrated in multi-unit firms than is the case for R&D performers as a whole and for the economy as a whole.

Top2003 firms have over 50,000 establishments associated with them in 2003. The number of establishments in each firm is very heterogeneous. The median number of establishments in these firms is 60, with the maximum number of establishments exceeding a thousand establishments. By comparison, the median number of establishments for R&D performers who are multi-units and the median number of establishments for multi-units in the total economy are both 3.[37]

As part of our interest in firm structure, we will also be looking at whether the R&D activities take place within stand-alone R&D laboratories or in production facilities. Scherer and Huh (1992) notes that U.S. Census data suggest that “more than half of all industrial R&D employees work outside free-standing laboratories or other central offices; that is, they are employed within plants that also produce goods for sale.” The prevalence of stand-alone laboratories may be related to the industry of the firm. National Research Council (2005, p. 23) notes that “Unlike traditional manufacturers, for which

[37] The median number of establishments by firm for the entire economy in 2003 is 1. That is, most firms are single-units.

most innovations grow out of dedicated R&D laboratories, more of the innovation in financial institutions comes from within operating units."

It is not possible to determine the exact location of the R&D expenditures for the R&D performing firms. Thus, it is not possible to distinguish between firms who perform their R&D at multiple locations from those firms who perform their R&D in a centralized location (unless they occur over multiple states). However, using the LBD we can provide some information about the prevalence of establishments that are dedicated to R&D. We define these dedicated establishments as those establishments that are classified as being in Scientific Research and Development Services (NAICS 5417, hereafter R&D Services).

In 2003, R&D performing firms where the firm as a whole is classified in R&D Services had 12.5 billion dollars in R&D total expenditures, which represents about 6% of total R&D expenditures. Thus, most R&D expenditures are taking place in firms that are not classified as dedicated to scientific R&D research. Just under 5% of R&D performers in 2003 have at least one establishment in this industry. However, more than 60% of the Top2003 firms have at least one establishment in R&D Services. Of these Top2003 firms with labs, about 17% have a single lab and about 25% have two to five labs.

4.4 Market Structure

We would like to have a measure of market concentration at the industry level that could be used to see whether top performing R&D firms are more likely to be in concentrated industries. As noted earlier, the industry codes used in 2003 have some issues. These issues were addressed for the 2004 via industry corrections to the top 500 companies in the SIRD. We will create our measure of market concentration by applying these corrected codes to our sample and then using existing measures of industry concentration. Given that we cannot match all of the R&D performing firms in 2003 with codes in 2004, this means that this information is only available for continuing firms. We will use the four-firm concentration data from the 2002 Economic Censuses applied at the 4-digit NAICS level as a starting point. Using this information, we will create a dummy variable for concentration and thus pool together information across industries which we would not otherwise be able to disclose. Industries are classified as highly concentrated if their four-firm concentration ratio is 80% or higher (this was the legal definition of highly concentrated before the Herfindahl-Hirschman Index became the standard measure). In future drafts of this paper, we will present results using this measure.

4.5 Summarizing Characteristics of R&D Performers and of Top2003 Firms

The preceding discussion has provided a detailed portrait of Top2003 firms relative to other R&D performing firms and to other firms more generally. In this section, we use regression analysis in an attempt to discover the significance of differences between Top2003 firms and other firms. We begin by examining the characteristics of R&D performing firms. We start with a simple set of regressions examining the relationship between the size of the R&D activity and firm characteristics. The regressions are for all R&D performers in 2003 and are sample weighted. All regressions

include four-digit NAICS industry controls.[38] We consider five specifications for the regressions where we start with the relationship between log R&D expenditures and each of four relevant characteristics and then one regression which includes all of the characteristics. The firm characteristics are size (as measured by total employment, firm age, scientists and engineer employment, and number of research laboratories).

Regression results are shown in Table 4. Consistent with our earlier results, we find that larger firms have greater total R&D expenditures (specification 1). This is consistent with what Akcigit and Kerr (2010) find for reference year 1997. They find that the mean R&D expenditures for R&D performing firms rises monotonically from $3.8 million for the smallest firms (100 employees or less) to $149 million for the largest firms (greater than 5,000 employees).

Turning to firm age, we find that very old firms spend more than very young firms, but that there is not a monotonic relationship between R&D expenditures and firm age.[39] When we control for firm size and firm age, we find some evidence that it is the youngest firms that have higher R&D expenditures.

The number of scientists and engineers is positively related to R&D expenditures, as is the number of research laboratories. However, the coefficient on the number of research laboratories is insignificant.

Together these firm characteristics explain a large portion of the variation in R&D expenditures. Controlling for all firm characteristics, we find an R-square of 0.74.[40] These regressions do not control for selection and therefore should be viewed only as a rough first approximation. Most small, young firms do not have R&D expenditures and thus are not part of our regressions. In future drafts of this paper, this regression will control for selection. We will also present results that incorporate the measure of market concentration.

5. Dynamics of Top R&D Performing Firms

We now turn our attention to the dynamic properties of Top2003 firms. We start by examining the persistence of Top2003 firms. Here we are interested in whether the firms have existed for long periods of time and whether they have been large R&D performers for long periods of time. We briefly consider the flip-side of persistence and examine whether the top 200 firms in 1976 exist in 2003 and are large R&D performers in 2003. Lastly, we examine the evolution of the Top2003 firms over time.

5.1 Persistence of Top2003 Firms

Top2003 firms are persistent relative to firms in the whole economy. By design, firms can be out of the SIRD sample and still be in existence. Therefore, we use the LBD to examine the persistence of Top2003 firms. As an initial exercise, we look for the first

[38] We also run the regression on the 4-digit NAICS industry controls alone and find an R-square of 0.33.
[39] The omitted age category is firms that are 10 years or younger.
[40] Running the regression controlling for all firm characteristics, but without including 4-digit NAICS industry controls yields an R-square of 0.61. Also, running the regression controlling for all firm characteristics (including 4-digit NAICS industry effects) except the number of scientists and engineers yields an R-square of 0.61.

appearance of the firm identifier in the LBD. Over half (57%) of Top2003 firms appear in the first year of the LBD, 1976, while only 20% of multi-unit firms in the whole economy in 2003 appear in the first year of the LBD. R&D performers are also persistent relative to firms in the whole economy; 30% of 2003 R&D performers and 14% of firms in the whole economy appear in the first year of the LBD.

The first appearance of the firm identifier in the LBD is not always a good measure of persistence because mergers and acquisitions can change firm identifiers over time. Therefore, we also look for the first appearance of any establishment of the firm in the first year the firm identifier appeared in the LBD. By this measure, 65% of Top2003 firms, 41% of multi-units, 46% of R&D performing firms, and only 17% of firms in the whole economy appear in the LBD in 1976.

Figure 6 shows the percent of Top2003 firms in existence by year using the following methods to define the firm's first year: (1) first year firm identifier appears in the SIRD; (2) first year firm identifier appears in the LBD; and (3) first year oldest establishment in the firm (in the first year the firm appear in the LBD) appears in the LBD. Many more firms are identified as alive using the match to the LBD. This figure demonstrates that it is necessary to match the SIRD to the LBD to look at the evolution of Top2003 firms.

5.2 Persistence of Top1976 Firms

While our focus in this paper is on the top 200 R&D performing firms in 2003, we take a moment here to consider the top R&D performing firms in 1976 (henceforth referred to as Top1976). In the previous subsection, we looked at the persistence of Top2003 through a view backwards over time. In this subsection, we look at the persistence of Top1976 firms through a view forwards over time. The start date in this section is 1976, since that is the year in which we know details about the firm (we can determine if the firm is in existence back to 1975). We find that 58% of the Top1976 are still in existence in 2003 according to the LBD and about 55% of Top1976 firms are in the 2003 SIRD. In terms of their R&D performance over time, over 30% of the Top1976 are still in the top 200 performing R&D ranking by 2003.

We are interested in comparing the Top1976 firms who managed to stay on top to those firms did not stay at top of R&D performers or who stopped existing altogether. In future drafts of this paper, we intend to examine these differences. As is always the case, we need to be especially careful about mergers and acquisitions when thinking about firms that are in the top ranking in one year but not in another year. For example, a number of Top2003 firms do not appear in the top 200 firms in 2007 simply because a merger or acquisition has caused a spurious break in the firm linkage over time.

5.3 Evolution of Top2003 Firms

We examine the evolution of Top2003 firms in this section. We are careful to use the LBD to determine whether or not a firm is in existence for a particular year. In our initial examination of the evolution of Top2003 firms, we focus on three broad types of firms: large continuers, growing continuers, and new firms. The first group refers to firms that were already large in terms of R&D activity at the start of our sample. The second

group refers to firms that were in existence at the start of our sample but that were not large R&D performers. Finally, the last group refers to firms that were not in existence at the start of our sample.

We start by examining these three groups focusing on their R&D activity. Since we are interested in the evolution of R&D activity over time, we create a measure that compares R&D activity in 2003 to R&D activity in 1976 (the ratio of real total R&D expenditures in 2003 to real total R&D expenditures in 1976). We define stable continuers as those firms whose ratio of R&D activity was relatively stable over time (specifically, as a ratio less than two). We define growing continuers as firms whose real R&D activity more than doubled over time. These measures compare the firm in 2003 to the firm in 1976 and thus account for firm-effects. However, we need to be careful in this interpretation since the firm itself may have changed through mergers and acquisitions over this time. Thus, we also use an alternate definition based upon absolute rankings within a cohort, that is, our measure of Top1976. Using this definition, cohort-stable firms are firms that are in the Top1976. Cohort-switching firms are firms that were in existence in 1976 but who were not in the Top1976. The last group of firms consists of firms born after 1976 where we can see their complete evolution from birth.

Recall that about 70% of the Top2003 firms can be considered continuers from 1976 (see section 4.2). Unfortunately only 40% of the Top2003 firms appear in the SIRD in 1976. That is, we will only be able to construct our measures of stable/growing and cohort-stable/cohort-switching continuers for slightly more than half of the relevant firms. Of these firms, one-third had real R&D total expenditures that were about the same magnitude in 2003 as in 1976, while the other two-thirds at least doubled their total R&D expenditures over this period. Using the cohort definition, one-quarter of the firms moved into the top cohort over time, while three-quarters of the firms were in the top cohort in both 1976 and 2003. In sum, of the Top2003 firms who were in existence in 1976 and for whom we can measure their R&D expenditures, most of them were in top cohort as well in 1976 and most of them held on to their position by more than doubling the amount of R&D expenditures over time. For those firms that were not in the top cohort in 1976, not surprisingly, their change in R&D expenditures was even more dramatic thus allowing them to push into the top cohort by 2003.

The last group consists of firms that entered after 1976, about 35% of Top2003 firms.[41] For this group, we can examine their complete evolution. We are particularly interested in whether these firms start as small firms as most firms in the economy do,[42] or whether they start as larger or even multi-unit firms. We can also examine their growth over time. If these firms are like those surviving firms in the rest of the economy (recall that we know these firms survive to 2003), we would expect that these firms will show rapid growth in size when they are young and that much of this growth will occur as an expansion of existing establishments rather than as adding more establishments (see Jarmin, Haltiwanger and Miranda (2010)). We find that over 70% of these firms entered as single unit firms and that they have a median employment of about 15 employees in

[41] We note again here that we have not yet carefully handled mergers, acquisitions, and spin-offs. We plan a more careful treatment of these cases in the next draft of this paper.
[42] Jarmin, Haltiwanger, and Miranda (2010, p. 10) find that "Nearly all firm births are small."

their birth year.[43] The birth year differs for these new firms, but five years out from their birth year, these firm have a median employment of close to 500 employees.

As noted earlier, most of the Top2003 firms have at least one establishment that could be considered a research lab based upon its industry classification (NAICS 5417). We are interested in whether these research establishments were with the firm from the start or whether they were establishments that were added later. If they were added later, we are also interested in whether they are de novo establishments or if they were acquired from another firm. In future drafts we will examine the share of the research labs were added to the firm after the firm was in existence. (Note that we will exclude all observations from this analysis where the research lab and the firm have the same initial date.) Further, we will look at the share of these labs that were acquired from another firm.

6. Conclusions and Future Research

We use three micro-level datasets to create a detailed picture of the top R&D performing firms in the U.S. We look at characteristics of these firms in 2003 and then follow the top 200 R&D performing firms back through time to describe both their persistence and their evolution. In all of our empirical analyses, we attempt to avoid areas where the underlying data are not as strong (such as in location and technology) and instead focus on areas where the data are of higher quality. We find that the top R&D performing firms are large, multi-unit, older firms. Many of the top R&D performing firms in 2003 are also among the top performing firms in 1976. In order to stay at the top, these firms had to more than double their R&D expenditures over the time period. Of the top R&D performing firms who were not top performing firms in 1976, some were in existence then and moved to the top by greatly expanding their R&D expenditures. Lastly, some of the top R&D performing firms in 2003 were not even in existence in 1976. Most of these entering firms started out as relatively small single units and then experienced rapid growth.

While we have found many interesting characteristics of the top R&D performing firms in our work, there are many areas where we could refine and improve our existing work. First and foremost, we have areas where we can improve our *measurement* of key firm characteristics. We would like to spend some time improving our handling of mergers and acquisitions. We have tried to be careful about making statements concerning births and deaths by relying heavily on the LBD to determine existence, but when working with firms, mergers and acquisitions can greatly complicate the analysis. We would also like to improve our planned measure of market concentration. We currently plan to first use a four-firm concentration ratio applied at the industry level. Acs and Audretsch (1987, p. 570) note that "Concentration ratios are a weak measure of market power at best, and in practice contain substantial error." We plan to start by also using a Herfindahl index to measure market power.

[43] We define a firm's birth year in the same way we define firm age as the first year the oldest establishment in the firm (at the first appearance of the firm ID in the LBD) appears in the LBD. We designate a firm as starting out as a single-unit if the firm has one establishment in its first year *and* that establishment is a single-unit.

One area of measurement we plan to explore in a future paper is defining innovative activity. In this paper, we have used total R&D expenditures as an indicator of innovative activity. We have mentioned in our paper many other measures used by researchers including company R&D expenditures, counts of innovations from trade journals, and patent counts. Each of these measures has strengths and weaknesses; some are simply not available for all of the firms in our sample (for example, the counts of innovations). Others we hope to explore in future papers. It seems especially important to consider using company R&D as an alternate measure of R&D activity. There are detractors of using patent data alone to measure innovations, so we would like to add patent counts as an additional piece of information.[44]

We also have many areas where we could expand on our existing work. It would be very interesting to take the flip side of our analysis and to find out how many of the fastest growing firms in the LBD had R&D expenditures. This work would complement the work in Akcigit and Kerr (2010) and some exploratory work that Jarmin and Miranda have done using the LBD. Another interesting avenue of research would be to push back the time period in which we can examine the evolution of our top performing firms by adding in the nonemployer universe. It may be that some of the firms transitioned at some point from a small one-person garage operation to a small employer. A much more ambitious idea would be to use the LEHD data to track employees as they move to different firms and perhaps start up their own new firms. Another more ambitious line of research would be to include some analysis of the outcomes of the R&D expenditures. Using Census microdata, we could measure the productivity growth at our top R&D performing firms and compare to the productivity growth of other firms. In summary, there are many avenues that could be explored using these rich datasets to further our understanding of the innovative processes in the U.S. economy.

[44] For example, Acs and Audretsch (1987) reference Shepherd (1979) in noting that patents are a "notoriously weak measure" of innovation since many never bear fruit and some are used simply to impede the innovations of others. On the other hand, Griliches (1990) notes many arguments in their favor.

References

Acs, Zoltan J. and David B. Audretsch. 1987. "Innovation Market Structure and Firm Size." *The Review of Economics and Statistics*, 69(4): 567-574.

Acs, Zoltan J. and David B. Audretsch. 1988. "Innovation in Large and Small Firms: An Empirical Analysis." *The American Economic Review*, 78(4): 678-90.

Adams, James. 1997. "The Structure of Firm R&D and the Factor Intensity of Production." Center for Economic Studies Discussion Paper CES-WP-97-15.

Adams, James and Adam Jaffe. 1994. "The Span of the Effect of R&D in the Firm and Industry." Center for Economic Studies Discussion Paper CES-WP-94-7.

Adams, James and Suzanne Peck. 1994. "A Guide to R&D Data at the Center for Economic Studies U.S. Bureau of the Census." Center for Economic Studies Discussion Paper CES-WP-94-9.

Akcigit, Ufuk and William R. Kerr. 2010. "Growth Through Heterogenous Innovations." Working paper, July 14, 2010.

Atrostic, B.K. 2008. "Measuring U.S. Innovative Activity: Business Data at the U.S. Census Bureau." *Journal of Technology Transfer*, 33:153-171.

Balasubramanian, Natarajan and Jagadeesh Sivadasan. Forthcoming. "What Happens when Firms Patent? New Evidence from U.S. Economic Census Data." *The Review of Economics and Statistics*.

Becker, Randy and Cheryl Grim. 2009. "Recovering Historical Manufacturing Microdata." *Center for Economic Studies, Annual Report 2008*: 29-31.

Cohen, Wesley M., Richard C. Levin, and David C. Mowery. 1987. "Firm Size and R&D Intensity: A Reexamination." *The Journal of Industrial Economics*, 35(4): 543-565.

Cohen, Welsey and Steven Klepper. 1996. "Firm Size and the Nature of Innovation within Industries." *The Review of Economics and Statistics*, 78(2): 232-43.

Fairman, Kristin, Lucia Foster, C.J. Krizan, and Ian Rucker. 2008. "An Analysis of Key Differences in Microdata: Results from the Business List Comparison Project." Proceedings of the American Statistical Associations, Business and Economics Statistics Section [CD-ROM], Alexandria, VA: American Statistical Association.

Foster, Lucia and Cheryl Grim. Forthcoming. "Creating the Top 200 R&D Performers Dataset." Center for Economic Studies Technical Report. Available to researchers with approved projects.

Griliches, Zvi. 1990. "Patent Statistics as Economic Indicators: A Survey." *Journal of Economic Literature*, 28: 1661-1707.

Griliches, Zvi. 1986. "Productivity, R&D and Basic Research at the Firm Level in the 1970s." *American Economic Review*, 76(1): 141-54.

Griliches, Zvi. 1980. "Returns to Research and Development Expenditures in the Private Sector." In *New Developments in Productivity Measurement*, ed. J.W. Kendrick and B. Vaccara, 419-61. Chicago and London: NBER/University of Chicago Press.

Jarmin, Ron S. and Javier Miranda. 2002. "The Longitudinal Business Database." Center for Economic Studies Discussion Paper CES-WP-02-17.

Jarmin, Ron S. John Haltiwanger, and Javier Miranda. 2010. "Who Creates Jobs? Small vs. Large vs. Young." Center for Economic Studies Discussion Paper CES-WP-10-17.

Kerr, William and Shihe Fu. 2008. "The Industry R&D Survey: Patent Database Link Project." *Journal of Technology Transfer*, 33(2): 173-186.

Levin, Richard C., Wesley M. Cohen, and David C. Mowery. 1985. "R&D Appropriability, Opportunity, and Market Structure: New Evidence on Some Schumpeterian Hypotheses." *The American Economic Review*, 75(2): 20-24.

Long, William and David Ravenscraft. 1993. "LBOs, Debt and R&D Intensity." Center for Economic Studies Discussion Paper CES-WP-93-3.

Mansfield, Edwin. 1981. "Composition of R&D Expenditures: Relationship to Size of Firm, Concentration, and Innovative Output." *The Review of Economics and Statistics*, 63(4): 610-15.

National Research Council. 2005. *Measuring Research and Development Expenditures in the U.S. Economy.* Panel on Research and Development Statistics at the National Science Foundation, ed. Brown, Lawrence D., Thomas J. Plewes, and Marisa A. Gerstein. Committee on National Statistics, Division of the Behavioral and Social Sciences and Education. Washington, D.C.: The National Academies Press.

National Science Foundation. 2009a. "U.S. Business R&D Expenditures Increase in 2007; Small Companies Performed 19% of Nation's Business R&D." NSF 09-316.

National Science Foundation. 2009b. "Research and Development in Industry: 2004." NSF 09-301.

National Science Foundation. 2007a. "Expenditures for U.S. Industrial R&D Continue to Increase in 2005; R&D Performance Geographically Concentrated." NSF 07-335.

National Science Foundation. 2007b. "Research and Development in Industry: 2003." NSF 07-314.

National Science Foundation. 1995. "Research and Development in Industry: 1992." NSF 95-324.

Nguyen, Sang and Edward Kokkelenberg. 1991. "Measuring Total Factor Productivity, Technical Change and the Rate of Returns to Research and Development." Center for Economic Studies Discussion Paper CES-WP-91-3.

Scherer, F.M. and Keun Huh. 1992. "R&D Reactions to High-Technology Import Competition." *The Review of Economics and Statistics*, 74(2): 202-12.

Scherer, F.M. 1992. "Schumpeter and Plausible Capitalism." *Journal of Economic Literature*, 30(2): 1416-33.

Schumpeter, Joseph A. 1942. *Capitalism, Socialism and Democracy*. New York: Harper.

Siegel, Donald and Frank Lichtenberg. 1989. "Using Linked Census R&D-LRD Data to Analyze the Effect of R&D Investment on Total Factor Productivity Growth." Center for Economic Studies Discussion Paper CES-WP-89-2.

Teece, David J. 1992. "Competition, Cooperation, and Innovation Organizational Arrangements for Regimes of Rapid Technological Progress." *Journal of Economic Behavior and Organization*, 18: 1-25. Collected and reprinted in *Essays in Technology Management Policy: Selected Papers of David J. Teece*.

Table 1. Summary Information on the SIRD Sample, 2003

Activity	**Number**
Frame Partition	
Companies known to conduct R&D	8,811
Companies known not to conduct R&D	74,318
Companies with unknown R&D activity	1,748,720
Sample Selection	
Certainty companies	10,014
Non-certainty companies[1]	21,484
Survey Forms Sent	
RD-1	3,039
RD-1A	28,882

Source: National Science Foundation (2007b).

Note:
[1] Non-certainty companies in the group of companies known not to conduct R&D in recent years are sampled using simple random sampling. Other non-certainty companies are sampled with probability proportional to size.

Table 2. Summary Statistics from the SIRD, 2003

	2003[1]	
	R&D Performers	**Top2003**
R&D Expenditures ($M)		
Total	200,724[3]	131,015
Mean	5	655
Median[2]	0.1	275
Standard Deviation	85	967
R&D Intensity (R&D Expenditures/Sales)		
Median[2]	0.03	0.12
Domestic Net Sales ($M)		
Total	5,748,522[3]	2,048,137
Mean	148	10,241
Median[2]	3	3,534
Standard Deviation	1,923	18,722
Domestic Employment		
Total	15,337,000[3]	4,983,308
Mean	404	24,917
Median[2]	28	11,272
Standard Deviation	4,437	40,778
Number of Scientists and Engineers		
Total	1,075,500[3]	597,313
Mean	30	2,987
Median[2]	2	1,336
Standard Deviation	404	4,378
N	37,843[3]	200

Source: Authors' calculations on the 2003 SIRD and National Science Foundation (2007b).

Notes:

[1] All statistics (including *N*) are weighted by SIRD sample weights.

[2] Medians shown are "fuzzy" medians calculated as the mean of all observations between the 45th and 55th percentiles.

[3] For disclosure reasons, the number shown in this cell is from the published SIRD statistics in National Science Foundation (2007b).

Table 3. Summary Statistics from the SIRD-LBD Matched Data, 2003

	2003			
	Whole Economy (LBD)[1]	**Whole Economy - MUs (LBD)[1]**	**R&D Performers (Match to LBD)[2]**	**TOP2003 (Match to LBD)[2,4]**
Firm Structure				
Multi-Units (%)	8	100	21	100
Employment at Multi-Units (%)	57	100	92	100
Firm Size (LBD Employment)				
Total	107,019,969	61,087,243	16,033,890	5,245,383
Mean	47	406	444	26,492
Median[3]	10	52	28	12,934
Standard Deviation	1,220	4,709	4,862	41,845
Firm Age				
Mean	13	19	16	23
Median[3]	11	25	16	27+
Standard Deviation	9	9	9	6
N	2,265,066	172,267	36,078	200

Source: Authors' calculations on the 2003 LBD and SIRD.

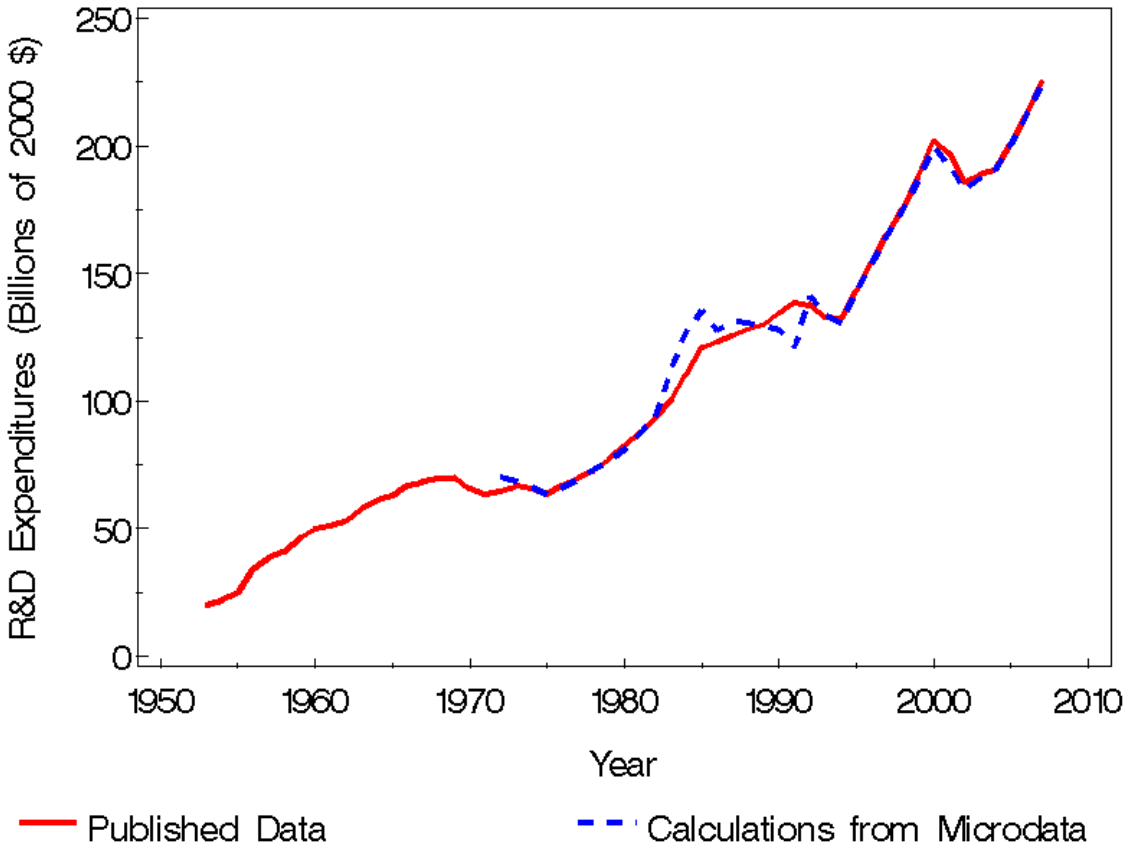
Notes:
[1] Whole economy data are limited to firms with at least five employees to be consistent with the SIRD sample frame.
[2] All statistics for R&D Performers and Top2003 firms (including *N*) are weighted by SIRD sample weights.
[3] Medians shown are "fuzzy" medians calculated as the mean of all observations between the 45th and 55th percentiles.
[4] To avoid potential disclosure of confidential data, some numbers shown in this column have been rounded up.

Table 4. Regression Results - R&D Expenditures and Firm Characteristics, 2003

Dependent Variable: Log R&D Expenditures

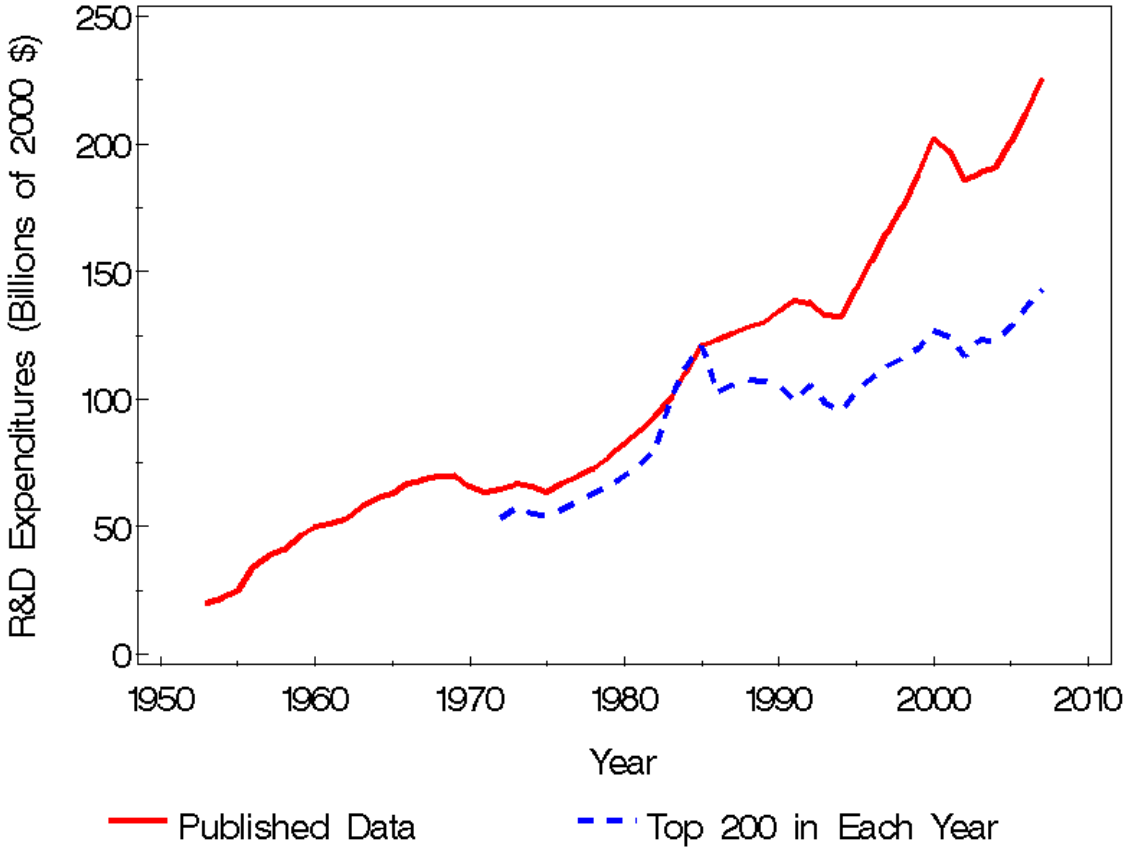
	(1)	(2)	(3)	(4)	(5)
Firm Employment (log)	0.95*				0.33*
	(0.01)				(0.02)
Firm Age 11-15		-0.37*			-0.33*
		(0.09)			(0.06)
Firm Age 16-20		0.55*			0.01
		(0.09)			(0.06)
Firm Age 20-26		0.55*			-0.21*
		(0.10)			(0.06)
Firm Age 27+		0.97*			-0.09
		(0.08)			(0.05)
S&E Employment (log)			1.26*		1.02*
			(0.01)		(0.02)
Number of Labs				0.07	-0.05
				(0.04)	(0.03)
Industry Effects	yes	yes	yes	yes	yes
R-square	0.62	0.36	0.74	0.35	0.74
N (weighted)	36,078	36,078	36,078	36,078	36,078

Notes: Regressions are run for firms with positive R&D expenditures in 2003 that can be matched to the LBD in 2003. Regressions are weighted by SIRD sample weight. The omitted firm age category is 0-10. The number of labs is the number of establishments in Scientific Research and Development Services (NAICS 5417). Standard errors are shown in parentheses. * denotes significance at the 5% level.



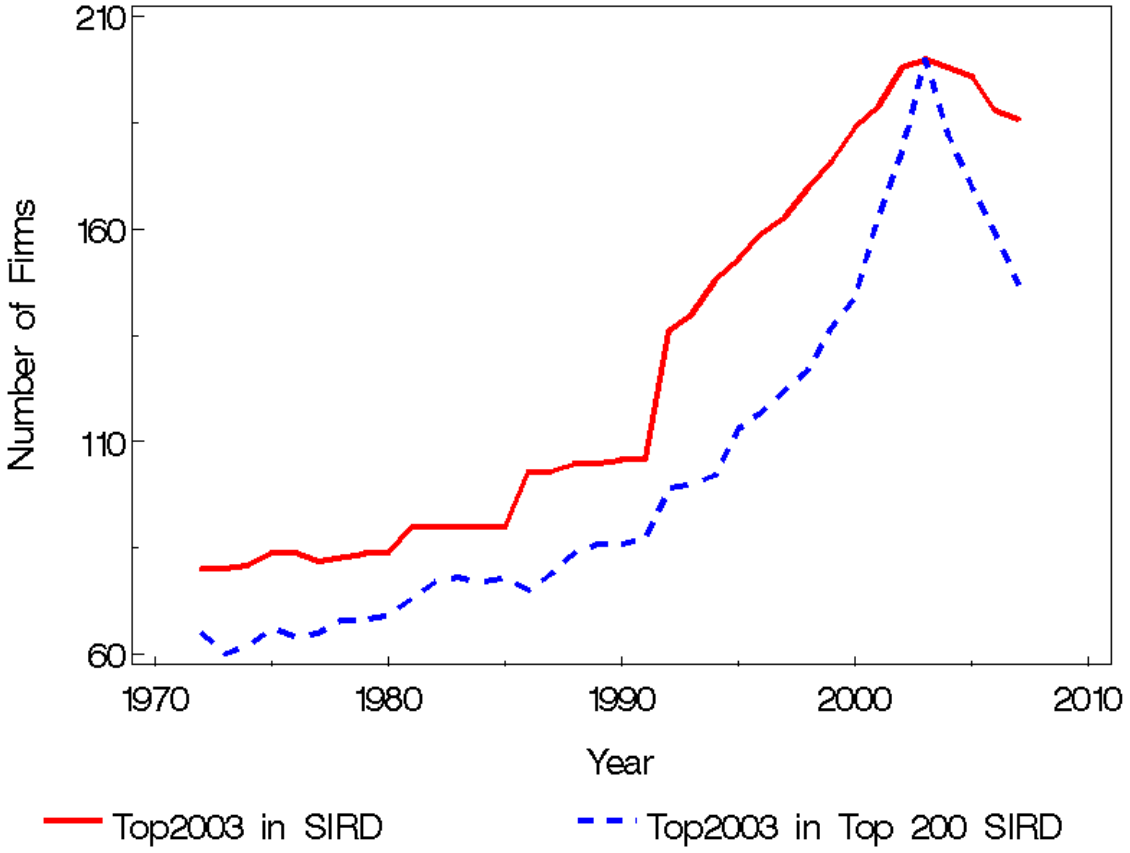
Sources: The 1953-2004 published data are from National Science Foundation (2009b), Table 2. The 2005 published data are from National Science Foundation (2007a), Table 1; the 2006 and 2007 published data are from National Science Foundation (2009a), Table 1. Calculations from microdata are from the authors' calculations on the 1972-2007 SIRD microdata files.

Figure 1. Total R&D Expenditures, 1953-2007



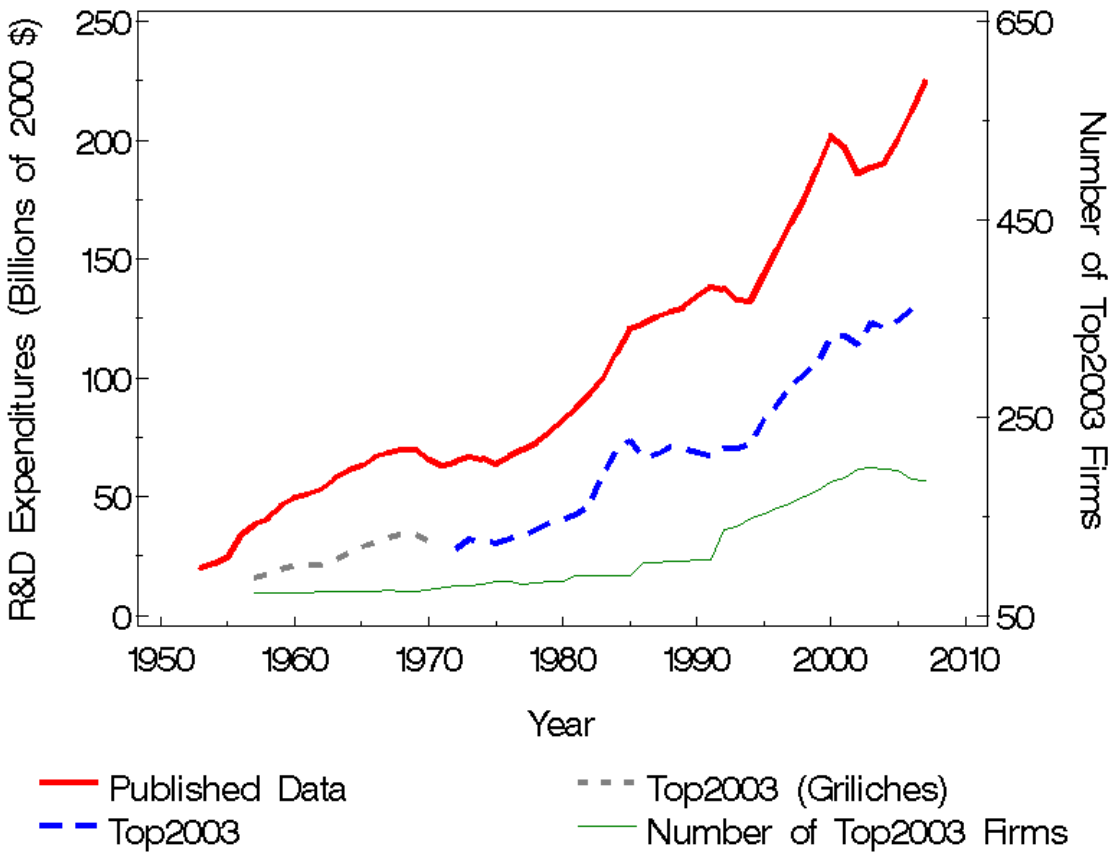
Sources: The 1953-2004 published data are from National Science Foundation (2009b), Table 2. The 2005 published data are from National Science Foundation (2007a), Table 1; the 2006 and 2007 published data are from National Science Foundation (2009a), Table 1. Top 200 calculations are from the authors' calculations on the 1972-2007 SIRD microdata files.

Figure 2. Total R&D Expenditures, All Firms and Top 200 R&D Performing Firms in Each Year



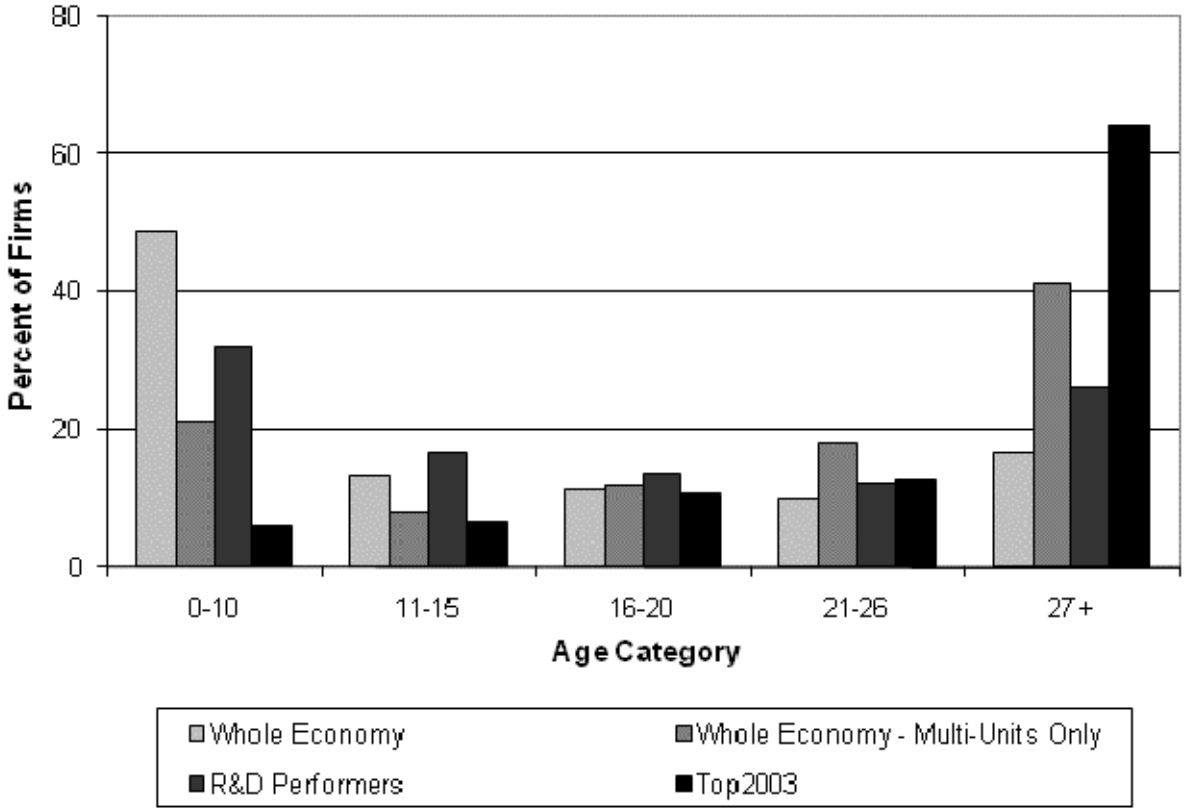
Source: Authors' calculations on the SIRD.

Figure 3. Number of Top2003 Firms, 1972-2007



Sources: The 1953-2004 published data are from National Science Foundation (2009b), Table 2. The 2005 published data are from National Science Foundation (2007a), Table 1; the 2006 and 2007 published data are from National Science Foundation (2009a), Table 1. Top2003 data are from the authors' calculations on the historical Griliches data and the 1972-2007 SIRD microdata files.

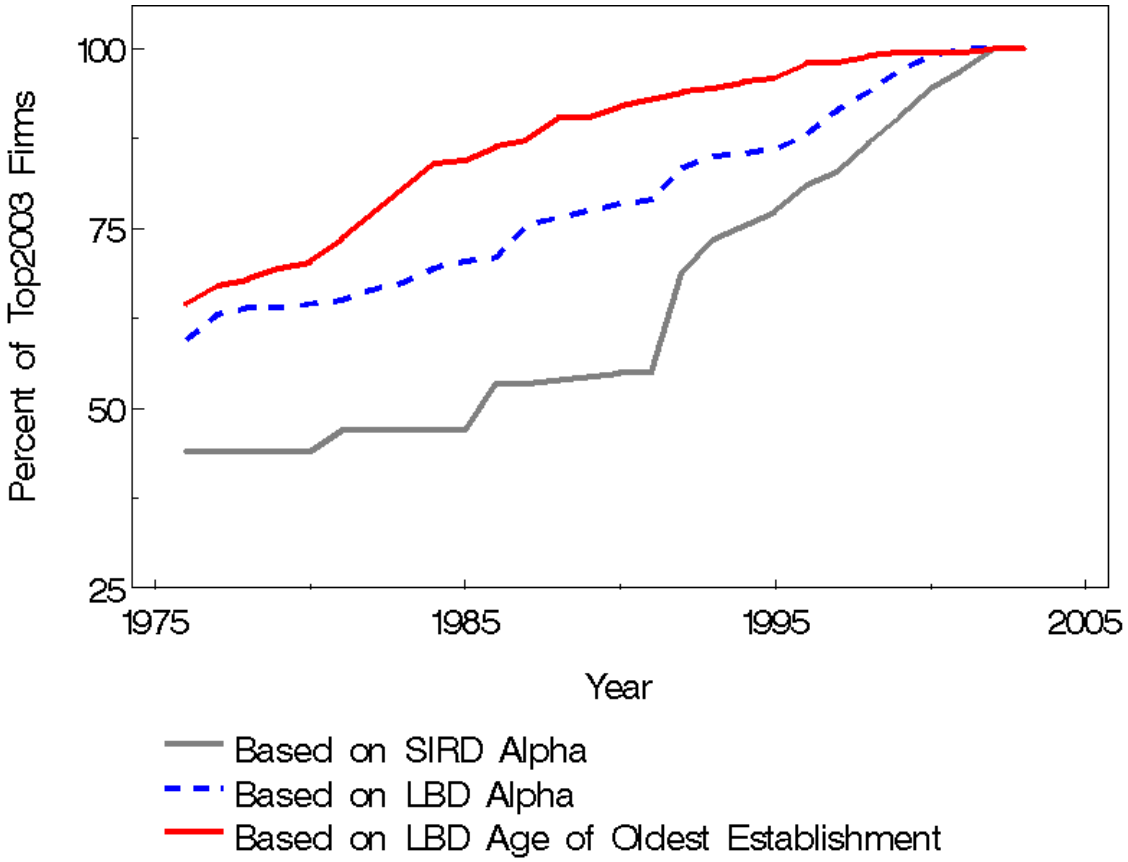
Figure 4. Total R&D Expenditures, All Firms and Top2003 Firms



Source: Authors' calculations on the LBD and SIRD.

Notes: The whole economy excludes firms with less than five employees for consistency with the SIRD sample frame. Firm age is calculated as the age of the oldest establishment in the firm as of 2003. Statistics for the positive R&D in 2003 group are calculated using SIRD sample weights.

Figure 5. Percent of Firms by Age Category, 2003



Source: Authors' calculations on the LBD and SIRD.

Figure 6. Percent of Top2003 Firms in Existence by Various Definitions, 1976-2003

Appendix

Total R&D expenditures appears in the 2003 form as question 4 reproduced below.

4 What was the cost of R&D performed within your company in the 50 United States and D.C.?

(Please report R&D performed for each source of funding.)

	2003								
	Federal funds (1)			Company and other (2)			Total (Columns (1)+(2)) (3)		
	$ Bil.	Mil.	Thou.	$ Bil.	Mil.	Thou.	$ Bil.	Mil.	Thou.
A. Basic research (Research for the advancement of scientific knowledge without specific immediate commercial objectives.)	304			305			306		
B. Applied research (Research directed primarily towards a specific commercial or practical objective.)	314			315			316		
C. Development (Activity translating research into new or improved goods, services or processes.)	324			325			326		
D. TOTAL *(Sum lines A through C)*	344			345			346		

The categories of items to be included in Total Expenditures are listed in the table below (From the SIRD instructions).

How to decide which expenditures to include as R&D costs

INCLUDE:	EXCLUDE:
• Wages, salaries, and related costs • Materials and supplies consumed • R&D depreciation • Cost of computer software used in R&D activities • Utilities, such as telephone, telex, electricity, water, and gas • Travel costs and professional dues • Property taxes and other taxes (except income taxes) incurred on account of the R&D organization or the facilities they use • Insurance expenses • Maintenance and repair, including maintenance of buildings and grounds • Company overhead including: personnel, accounting, procurement and inventory, and salaries of research executives not on the payroll of the R&D organization	• R&D from acquired companies prior to acquisition (in-process R&D) • Capital expenditures • Test and evaluation once a prototype becomes a production model • Patent expenses • Income taxes and interest