Volume Title: R & D, Patents, and Productivity

Volume Author/Editor: Zvi Griliches, ed.

Volume Publisher: University of Chicago Press

Volume ISBN: 0-226-30884-7

Volume URL: http://www.nber.org/books/gril84-1

Publication Date: 1984

Chapter Title: Patents and R&D at the Firm Level: A First Look

Chapter Author: Ariel Pakes, Zvi Griliches

Chapter URL: http://www.nber.org/chapters/c10044

Chapter pages in book: (p. 55 - 72)

3 Patents and R & D at the Firm Level: A First Look

Ariel Pakes and Zvi Griliches

3.1 Introduction

This paper is the first report from a more extensive study of knowledge-producing activities in American industry initiated by the National Bureau of Economic Research. Perhaps the most serious task facing empirical work in the area of "technological change" and "invention and innovation" is the construction and interpretation of measures (indices) of advances in knowledge.[1] If one defines K as the level of economically valuable technological knowledge, and $\dot{K} = dK/dt$ as the net accretion to it per unit of time, then the first task of our research program is to evaluate the usefulness of several indicators of $\dot{K}$, focusing particularly on patents and the value of the firm, variables which have yet to receive the attention that we think might be warranted in this context.[2]

The basic structure of our project is illustrated succinctly by the path analysis diagram in figure 3.1. In that diagram $\dot{K}$ is a central unobservable which, together with the observables, the X's, and the disturbances, the v's, determines the magnitude of several interrelated indicators of invention and innovation, the Z's. The latter include the stock market value of the firm, the productivity of traditional factors of production, and invest-

Ariel Pakes is a lecturer in the Department of Economics in the Hebrew University of Jerusalem, and a faculty research fellow of the National Bureau of Economic Research. Zvi Griliches is professor of economics at Harvard University, and program director, Productivity and Technical Change, at the National Bureau of Economic Research.

1. For a thoughtful discussion of this point, see Kuznets (1962).

2. Most of the previous work on patents is either quite ancient or inconclusive. Professional opinion has not really progressed much past the disagreement about the utility of patent statistics reflected in the discussions between Kuznets, Sanders, and Schmookler (Nelson 1962). The most recent review of the literature and independent contribution is found in Taylor and Silberston (1978). The papers that come closest to the topics treated here are Scherer (1965) and Comanor and Scherer (1969).

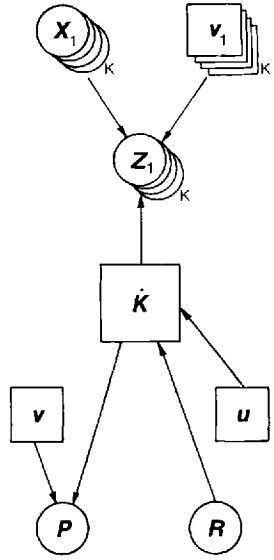

Fig. 3.1 A simplified path analysis diagram of the overall model.

ment expenditures on traditional capital goods. We report on an investigation of the lower half of this diagram in this paper. We see in figure 3.1 that $\dot{K}$ is produced by a knowledge production function (KPF) which translates past research expenditures, R, and a disturbance term, u, into inventions. The disturbance term reflects the combined effect of other nonformal R & D inputs and the inherent randomness in the production of inventions. Patents, P, are an imperfect indicator of the number of new inventions, with v_0 representing the noise in the relationship between P and $\dot{K}$. It is clear from the figure that the patent equation, the equation connecting patents to past research expenditures, combines the properties of both the KPF and the indicator function relating to P and $\dot{K}$. Without additional indicators of $\dot{K}$ one cannot separate the two types of effects. For example, both u and v_0 enter the relationship between R and P, but only u affects the Z's. In the context of a larger model, one could separate out the effects of u from v_0 by calculating the effect of the residual in the patent equation on the Z's, but this cannot be done from the patent equation alone.

We have made several simplifications in drawing and discussing this diagram. For example, the relationship between K and $\dot{K}$ should be defined explicitly to allow for the possibility of decay in the private value of knowledge. $\dot{K}$ may be determined by the absolute level of K as well as by past investments in research resources. If, as in likely, the u's are correlated over time, then one would expect any realization of u to feed

back into the demand for research resources. Moreover, conditions (economic, technological, and legal) should be specified under which the benefits from applying for a patent outweigh the costs of the patenting process, adding thereby more structure to the relationship between P and $\dot{K}$.[3] Figure 3.1 does, however, provide an overview of our project and is sufficiently precise for the discussion of the two issues on which this paper will concentrate: (1) the "quality" of patent counts as an indicator of knowledge increments, and (2) the time shape of the lag between research expenditures and patentable results.

The recent computerization of the U. S. Patent Office's data base has made it possible for the first time to follow the patenting behavior of a large cross section of firms over a significant time interval. This makes patent counts an easily accessible, perhaps the most easily accessible, indicator of the number of inventions made by a firm. Moreover, patents are a quantitative and rather direct indicator of invention; an indicator not contaminated by many of the X's which also affect the Z's. However, the patent measure does have several problems, the major ones being that not all new innovations are patented and that patents differ in their economic impact. These considerations have led to doubts about the "quality" of patent counts as an indicator of knowledge increments (see the literature cited in note 2). We attempt to respond to such concerns by first presenting a more precise description of the patent equation in section 3.2 and then reporting in section 3.3 on one particular measure of the "quality" of patent statistics.

Patent counts have another advantage over other indicators of knowledge production. Patents are applied for at an intermediate stage in the process of transforming research input into benefits from knowledge output. They can be used, therefore, to separate the lags that occur in that process into two parts: one which produces patents from current and past research investments, and another which transforms patents, with the possible addition of more research expenditures, into benefits. Such a breakdown should allow us to estimate more precisely the overall lag

3. Such a theory, we think, would be based on the underlying notion of a research project whose success depends stochastically on both the amount of resources devoted to it and the amount of time that such resources have been deployed. Each technical success is associated with an expectation of the ultimate economic value of a patent to the inventor or the employer. If this expectation exceeds a certain minimum, the cost of patenting, a patent will be applied for. That is, the number of patents applied for is a count of the number of successful projects (inventions) with the economic value of a patent exceeding a minimal threshold level. If the distribution of the expected value of patenting successful projects remains stable, and if the level of current and past R & D expenditures shifts the probability that projects will be technically successful, an increase in the number of patents can be taken as an indicator of an upward shift in the distribution of $\dot{K}$. Whether the relationship is proportional will depend on the shape of the assumed distributions and the nature of the underlying shifts in them. What we are dealing with here is at best a very crude reduced-form-type equation whose theoretical underpinnings still remain to be worked out. But one has to start someplace.

structure, a structure which has confounded and confused previous empirical work in this area.[4] Section 3.4 presents our first-round estimates of the distributed lag between research expenditures and patentable results.

The data used in this study are at the firm level and are based on a merger of the information provided in the Standard & Poor's (1980) Compustat file (based on the 10-K firm reports to the SEC) and patent data tabulated by the Office of Technology Assessments and Forecasts of the U.S. Patent Office. These data and the particular sample chosen are described in greater detail in Appendix A. Most of the work reported here is based on the patenting behavior of 121 firms during 1968–75.

3.2 The Model

We report in Appendix B a preliminary investigation into the functional form of the relationship between patents and past R & D expenditures. That analysis supports a rather simple patent equation: the logarithm of patents (p) as a function of a time trend (t), current and five lagged values of the logarithm of research expenditures (r), and a set of firm-specific dummy variables. In this section we provide an interpretation of this patent equation in terms of a simple model relating past r to the logarithm of current knowledge increments ($\dot{k}$), and $\dot{k}$ to p.

Consider first the transformation function from r to $\dot{k}$ or the KPF. Assuming it to be of the Cobb-Douglas form but allowing for firm constants and a time trend, we have:

$$\dot{k}_{i,t} = a_i + bt + \sum_{\tau=0}^{5} \theta_\tau r_{i,t-\tau} + u_{i,t}, \tag{1}$$

where $u_{i,t}$ is an independent and identically distributed disturbance which is not correlated with r and represents randomness in the KPF. The a_i represent firm-specific differences in the private productivity of research effort caused by either variation in appropriability environments, opportunities, or differences in managerial ability. Such differences will, in general, be transmitted to differences in research expenditures; firms with more productive research departments investing more in research. Thus, the a_i have two roles in the subsequent analysis. First, they cause differences in $\dot{k}$, and this should be considered in an analysis of the determinants of the variance in p. Second, their correlation with the $r_{t-\tau}$ must be accounted for in any attempt to estimate the θ_τ or else the coefficient estimates will be a combination of the effect of the $r_{t-\tau}$ on $\dot{k}$ (the θ_τ) and the effect of a_i or r. To be more explicit about the latter point,

4. See, for example, how two different assumptions about the lag structure lead to very different calculations of the private rate of return to research expenditures from the NSF-Griliches data: Griliches (1980b) versus Pakes and Schankerman (this volume).

we simply project the a_i on all in-sample research expenditures. Since the a_i are constant over time they should only be correlated with the means of the research variables. We can write, therefore,

$$a_i = \sum_{\tau=0}^{5} \psi_\tau r_{i,\cdot-\tau} + u_{i\cdot}, \tag{2}$$

where

$$r_{i\cdot 0} = T^{-1} \sum_{t=1}^{T} r_{it},\ r_{i\cdot -1} = T^{-1} \sum_{t=0}^{T-1} r_{it-1},\ \text{etc.},$$

and u_i is by construction uncorrelated with all in-sample research variables.[5]

Patents are our indicator of knowledge increments. If one allows for a time trend in the relationship between p and $\dot{k}$, that relationship is written as:

$$p_{i,t} = dt + \beta \dot{k}_{i,t} + v^*_{i,t}, \tag{3}$$

where v^* is uncorrelated with $\dot{k}$ and t by construction.

Equation (3) should be interpreted as a reduced form from the appropriate patenting model. In that reduced form, β is the elasticity of patents with respect to knowledge increments, and d is a measure of the trend in factors determining the propensity to patent. On the other hand, $v^*_{i,t}$ is that part of the (detrended) variance in patents which cannot be accounted for by (detrended) movements in knowledge increments; that is, variance in $v^*_{i,t}$ is "noise" in the patent measure. To facilitate interpretation we will make two assumptions on $v^*_{i,t}$. First, we let $v^*_{i,t}$ be composed of a firm-specific component, v_i, which reflects differences among firms in their *average* propensity to patent, and a second, independent, identically distributed disturbance, $v_{i,t}$, reflecting the *variations* (around a trend) in the propensity to patent of a given firm over time. Thus, $v^*_{i,t} = v_i + v_{i,t}$. Second, since $v^*_{i,t}$ is uncorrelated with $\dot{k}_{i,t}$ (by choice of β), we shall also assume that its determinants, v_i and $v_{i,t}$, are each uncorrelated with the determinants of $\dot{k}$ (the r's and u's) given by equations (1) and (2).[6]

5. In econometric terminology, the model we are working with is a variant of the partial transmission model of Mundlak and Hoch (1965). The unobservable portion of the KPF, which is transmitted to the research demand equation is assumed to remain constant over time. This assumption, plus the nature of the panel, will allow us to use single equation estimation techniques to estimate parameters of the patent production function. A more precise discussion of the econometric techniques underlying the estimation procedures to be used in this paper is found in Mundlak (1978) and Pakes (1978, chap. 3).

6. The first assumption allows us to provide standard errors for our estimates of the regression coefficients. The second is a rather strong assumption. We are assuming that randomness in the KPF, above or below average success in converting research expenditures into knowledge increments, does not influence the patenting decision, that the two sources of randomness are distinct and independent. We need this assumption to make the interpretations that follow.

Substituting (1) and (2) into (3), we can now provide an interpretation to the equation preferred in our analysis of functional form, that is to the equation:

$$p_{i,t} = \alpha + \gamma t + \sum_{\tau=0}^{5} w_\tau r_{i,t-\tau} + \sum_{\tau=0}^{5} \phi_\tau r_{i\cdot,-\tau} + \eta_i + \epsilon_{i,t}, \tag{4}$$

where

$$w_\tau = \beta\theta_\tau, \quad \gamma = \beta b + d, \quad \phi = \beta\psi,$$

$$\eta_i = v_i + \beta u_i, \text{ and } \epsilon_{i,t} = v_{i,t} + \beta u_{i,t}$$

for

$$i = 1, \ldots, N \text{ and } t = 1, \ldots, T.$$

The first point to note from equation (4) is that though one cannot estimate the elasticities of knowledge increments with respect to research resources, the θ_τ, one can investigate the form of the distributed lag connecting $\dot{k}$ and r, since $w_\tau/\Sigma w = \theta_\tau/\Sigma\theta$. The sum of the estimated lag coefficients, $w^* = \Sigma_{\tau=0}^{5} w_\tau$, estimates the product of the degree of economies of scale in the KPF, $\Sigma_{\tau=0}^{5} \theta_\tau$, and the elasticity of patents with respect to knowledge increments (β). These two parameters can be identified separately only in a larger model which includes additional indicators of the benefits from knowledge-producing activities (see section 3.1).

Recall that the various variance components which combine to form the disturbance term is (4) are mutually uncorrelated. It follows that $\mathrm{Var}(\eta_i + \epsilon_{i,t}) = \sigma^2$, the variance of the total disturbance in the patent equation, is greater than $\mathrm{Var}(v_i + v_{i,t})$, the variance of the noise in patents as an indicator of $\dot{k}$. It also implies that, temporarily ignoring the time trend in the patent indicator equation (assuming $d = 0$), the ratio of σ^2 to the total variance in the logarithm of patents ($1 - \bar{R}^2$) provides an upper bound for the noise-to-total-variance ratio in the patent measure. The upper bound will be called λ^{uT}, and its complement, the relevant $\bar{R}^2$ measure, is a measure of the quality of patents as an indicator of knowledge increments. If, instead of assuming $d = 0$, we assume $b = 0$, that is, the entire trend effect is caused by differences in the average propensity to patent over time, then one can derive an analogous measure of λ^{uT} for detrended patents by filtering out time from both the patent and the R & D variables. In practice, the two measures of λ^{uT} were always almost identical. In section 3.3 we also present the comparable information on the noise-to-total-variance ratio in the between firm variance in patents (i.e., in the variance of $p_{i\cdot} - p\ldots$), labeled λ^{uB}, and in the within firm variance in patents (the variance in $p_{it} - p_{i\cdot}$), λ^{uW}. The latter two statistics provide some indication of the usefulness of patent counts as an indicator of knowledge increments for studies of invention and innovation that

focus either on cross-section differences in the production of knowledge between firms or on the within firm fluctuations over time.

3.3 Measures of the Quality of the Patent Variable

Table 3.1 presents estimates of $1 - \lambda^{uT}$, $1 - \lambda^{uW}$, $1 - \lambda^{uB}$, the lower bounds to the systematic-to-total-variance ratios, σ_ϵ^2 and σ_η^2, and some relevant sample moments for each of the seven industries in our data (rows 0 through 6), all firms in our sample (row 7) and firms in the industries defined by rows 1 through 6 (row 8). The latter sample concentrates on firms in research-intensive industries.

Starting with the measures of $1 - \lambda^{uT}$ in the separate industries, it is clear, even from our simplistic model, that much of the patent variance is systematic, providing a good indicator of the underlying variance in $\dot{k}$. For the seven industries in our sample, about 85 percent of the variance in p is associated with variance in r, and in some industries, notably scientific instruments and office, computing, and accounting machinery, the lower bound of the systematic to total variance in patents is closer to .95.

These estimates hide, however, some relevant information. Moving to column (2), we are clearly far less certain of whether changes over time in p within any given firm reflect systematic changes in knowledge production by that firm. In the within firm calculations it mattered whether or not we first filtered out time trends from p and r. Therefore, the numbers in parentheses beside column (2) refer to systematic-to-total-variance ratios in detrended patents. Averaging over the seven industries, we find that the lower bound $(1 - \lambda^{uW})$ is only around 20–25 percent, though it does reach 50 percent in office, computing, and accounting machinery. Without the larger model alluded to in section 3.1, one cannot really tell whether the smaller systematic-to-total-variance ratios in the "within" data reflect true randomness in the knowledge production function (small differences in research expenditures over time within a given firm having very sporadic effects on the production of inventions in particular years) or whether they arise because firms decide to patent different proportions of their inventions in different years.

Two more points should be noted about the results for the separate industries. Column (6) shows that over 90 percent of the total variance in p is between firm variance. As a result $1 - \lambda^{uB}$ is very close to $1 - \lambda^{uT}$. Second, though σ_ϵ^2 does not vary too much between the sample industries, σ_η^2 varies a lot, being much larger in the less homogeneous industries (rows 0, 1, and 3). This is likely to reflect greater differences in the average propensities to patent in those industries.

Looking at the samples which aggregate the various industries (rows 7 and 8), we find that λ^{uW} actually decreases after pooling different industry samples. This implies that, at least in our sample, the elasticity of

Table 3.1 **Lower Bounds to the Systematic Variance Ratio in p and Some Sample Moments for the Seven Industries**

Industry Description[a]	Lower Bound to the Systematic Variance ratio[b]							Ratio of Within to Total Variance in p		Ratio of Within to Total Variance in r		
	$1-\lambda^{uT}$ total	$1-\lambda^{uW}$ within		$1-\lambda^{uB}$ between	σ_ϵ^{2c}	σ_η^{2d}	Variance in p		Variance in r		Firms N	Observations NT
	(1)	(2)		(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
0 Other manufacturing	.74	.19	(.16)	.77	0.17	.050	2.71	0.08	1.83	0.04	41	328
1 Industry 28 except 283 (chemicals & allied products except drugs & medicines)	.82	.13	(.11)	.86	0.10	0.26	2.02	0.06	1.16	0.02	19	152
2 Industry 283 (drugs & medicines)	.80	.33	(.22)	.85	0.07	0.14	0.85	0.10	0.56	0.03	19	152
3 Industry 35 except 357 (machinery except office, computing & accounting)	.82	.11	(.06)	.87	0.16	0.32	2.97	0.07	2.17	0.04	13	104
4 Industry 357 (office, computing & accounting)	.84	.52	(.50)	.96	0.16	0.09	3.78	0.11	2.06	0.11	10	80
5 Industry 366/367 (electronic components & communications)	.51	.46	(.42)	.89	0.07	0.07	0.83	0.13	1.57	0.03	8	64
6 Industry 38 (professional & scientific instruments)	.95	.28	(.06)	.97	0.08	0.05	2.55	0.04	2.21	0.04	11	88
7 Total sample	.66	.33	(.23)	.69	0.14	0.66	2.41	0.07	1.72	0.04	121	968
8 Firms in research-intensive industries	.73	.33	(.22)	.76	0.12	0.48	2.20	0.07	1.61	0.04	80	640

[a]Two- and three-digit industry identification numbers refer to SIC codes.

[b]The numbers in parentheses refer to lower bounds to the systematic variance ratio in patents after a time trend has been filtered out of both p and r: p − log patents; r − log R & D.

[c]Calculated as the variance of estimate from an OLS regression of $p_{i,t}$ on current and five consecutive lagged years of $r_{i,t}$ and firm-specific constants.

[d]Calculated as the variance of estimate from an OLS regression of equation (4) in the text minus the estimate of σ_ϵ^2.

patents with respect to knowledge increments (β) and the response of k to current and past $r(\theta_\tau)$ do not vary much between the industries aggregated; a result which will be confirmed in section 3.4.

3.4 Coefficient Estimates

Table 3.2 presents the estimates of the w_τ and the coefficient of the trend term based on data from all of the 121 firms and estimates based on two subsamples: firms in research-intensive industries and "other manufacturing" firms. Row 10 presents the estimates value of the F statistic for the null hypothesis that these coefficients do not differ between the industries aggregated. The test statistics indicate that, after we allow for a separate trend and intercept for the drug industry (row 9), our sample cannot really pick up any additional interindustry differences in coefficients.[7]

Turning to the coefficient of the trend term, that coefficient was negative, and significantly so, for all industries except for the drug industry. This result has two alternative explanations, and they cannot be separated out without the larger model alluded to in section 3.1. First, the negative trend is consistent with impressionistic evidence on the declining propensity to patent in U.S. manufacturing. The drug industry is indeed an exception since, during the period concerned, there occurred both a relaxation in the Patent Office's acceptance procedures regarding patents on natural substances and significant changes in regulatory conditions facing that industry.[8] The same result, however, could have been caused by a secular decline in the private productivity of research resources, a hypothesis which is consistent with the observed negative growth rate of employment of R & D scientists and engineers during the period considered.[9]

The individual coefficients are not estimated very precisely. The sum of the lags, w^*, is estimated with a fair amount of precision and equals about .60 with a standard error of 0.08. If one ignores the fact that some of the estimated lag coefficients are negative and computes a "mean lag," it equals about 1.6 years for the all-firm sample. Unless substantial R & D is done on projects after patents are applied for, this should approximately equal the mean R & D project gestation lag, the lag between project

7. The possible exception here is the drug industry. When that industry was dropped from the first two samples, the observed values of the F test dropped significantly to 1.37 and 1.67, respectively. Still the estimated coefficients for the drug industry were not very different from those of the other industries in the sample, except for the trend coefficient.

8. For a description of the effect of these events, see Temin (1979).

9. See Griliches (1980a) for a similar finding on aggregate data.

Table 3.2 **Distributed Lag Estimates**[a]

Variables	All Firms (1)	Firms in Research-Intensive Industries (2)	Other Manufacturing Firms (3)
1. r_0	.56	.52	.62
	(.07)	(.10)	(.14)
2. r_{-1}	−.10	−.01	−.22
	(.09)	(.12)	(.16)
3. r_{-2}	.05	.08	−.02
	(.09)	(.12)	(.16)
4. r_{-3}	−.04	−.21	.13
	(.09)	(.13)	(.15)
5. r_{-4}	−.05	−.01	−.08
	(.10)	(.15)	(.16)
6. r_{-5}	.19	.25	.13
	(.08)	(.11)	(.14)
7. Sum (w^*)	.61	.62	.61
	(.08)	(.09)	(.04)
8. t	−.04	−.05	−.03
	(.007)	(.008)	(.012)
9. t_{drugs}	.07	.07	—
	(.10)	(.01)	
10. F aggregation (critical values, 1%, 5%)	1.54 (1.39, 1.58)	2.08 (1.45, 1.69)	—
11. Degrees of freedom	837	550	279

[a]Standard errors are in parentheses below coefficient estimates.

inception and project completion. The scattered empirical evidence on gestation lags indicates that this is indeed the case.[10]

Still, the estimated form of the lag is rather disturbing. There are large, significant, postive coefficients in the first and last years and very little effect of interim R & D on patent applications. Though the current year's coefficient could indicate the presence of simultaneous equations bias, that is not really a necessary implication of the results. The R & D project level data cited above do point to a gestation lag highly skewed with large early year coefficients, and any minor misspecification in the model could push all this effect into the coefficient of r_0. The coefficient estimate which is perhaps more disturbing is that of the last year since it could be indicating the presence of a "truncation" problem in our distributed lag

10. Sources of project level data are Wagner (1968) and Rapoport (1971). This evidence is summarized in terms of mean gestation lags in Pakes and Schankerman (this volume). The average of the mean gestation lags presented in the latter paper was 1.34 years.

estimates. That is, the coefficient of the fifth year could be proxying for a series of small effects of the more basic research done six years ago or earlier.[11] These estimates of the form of the lag should be treated with caution, both because of the possible truncation problem and because they are not really consistent with our prior beliefs about the form of this lag structure.

3.5 Conclusions and Extensions

Our first look at the patent equation suggests the following conclusions. First, the data were quite clear on the form of that equation; log-log with (correlated) firm effects and a time trend being preferred over alternatives. Second, our major positive finding is given by the $1 - \lambda^{uB}$ estimates presented in table 3.1. They show that patents are a good indicator of between firm differences in advances of knowledge. Since the between firm component dominates the total variance in patents, a similar comment also applies to the total variance. If this result changes at all in the more sophisticated models we are beginning to estimate, it is only likely to improve. Use of a longer series of past R & D expenditures can only increase the fit of the patent equation, and adding another indicator of benefits will separate out the effect of randomness in the KPF, the u, from the effect of noise in the patent measure, the v^*, allowing us to narrow the bound further.

The rest of our results are not as heartening. While a part of the within firm variance in patents is related to the variance in R & D expenditures, a significant portion (about 75 percent) is not. At this stage we cannot tell whether the fault lies in the patent measure (the variance in v^*), in randomness in the KPF (the variance in u), or in simple errors of measurement in both p and r. Most of the coefficients, except for trend, were not estimated very precisely. This is a result of two factors: First, only the within firm variance in p and r can be used to estimate w_τ, and this variance is a small part of the total variance in these variables (see table 3.1). The second factor leading to imprecise estimates is the small sample size (maximum $T = 8$; $N = 121$). We can and will increase our sample significantly in the future by not insisting that firms had to have reported R & D expenditures before 1972 (see appendix A). Including such firms will force us, however, to use only a few lagged terms of r or assume a specific functional form for the distributed lag between patents and R & D expenditures, even though we have yet to acquire much information on the shape of this distribution. Because our estimates indicate that even with five lagged R & D terms we still may have a truncation problem, we have been developing a technique for estimating

11. See Griliches and Pakes (1980) for further discussion of such problems.

distributed lags in panel data when the time series on the independent variable is short. We are also investigating the impact of other sources of bias in the estimated coefficients, in particular the effect of measurement errors in the R & D variables. Finally, once an appropriate specification for the patent equation has been determined, we will combine it with the other equations in our model in the hope of providing a fuller understanding of the process of invention and innovation in American industry.

In short, a great deal of work remains to be done, but we have made a start. It is already clear that something systematic and related to knowledge-producing activities is being measured by patents and that they are, therefore, very much worthy of further study.

Appendix A

Data Sources, Sample and Variable Definitions, and Sample Characteristics

The data base used in this preliminary round is neither complete nor representative. We have tried to gather from published sources as large a sample of firms as possible covering 1963–77. The main selection variable is R & D. Until recently (1972 and later) most firms did not report their R & D expenditures publicly. The firms that did report R & D expenditures reported company-financed R & D expenditures, and those numbers are recorded on the Standard and Poor's Compustat tape, which served as a major source of our data.[12] An earlier study by Nadiri and Bitros (1980) had used both the Compustat tape and a mail survey to fill in some of the gaps on this tape to construct time series of R & D for 114 firms during 1963–72. Starting with a later edition of the tape, we found 146 firms with no more than three years of R & D data missing during 1963–75. Combining it with the Nadiri and Bitros sample yielded an unduplicated total of 172 firms. Fifteen firms were eliminated from this total either because they were foreign, had undergone large mergers, or had other unreconcilable jumps in their data. This left a total of 157 firms which constitute the data on which a number of recent NBER studies have been based.[13] Based on preliminary experimentation (see appendix B), the sample for this paper was further restricted to firms that had data (did not undergo any major reorganization) throughout the whole period

12. Only company-financed R & D ought to lead to patents since government R & D contracts most often include clauses which put the output of government-funded projects in the public domain.

13. For further description of this data base, see Bound and Hall, 1980. A much larger sample is possible to construct if one is willing to restrict oneself to post-1972 data. See Bound et al. (this volume).

($N = 144$) and had an R & D program of more than minimal size (R & D $\geq$ \$0.5 million) in any one year ($N = 121$). What we have done, then, is to expand the Nadiri and Bitros sample slightly, update it to 1977, *and* add patent data to it.[14]

The patent data were supplied to us by the Office of Technology Assessment and Forecasts (OTAF) of the U.S. Patent Office. They are based on a tape of all patents granted during 1969–78. These data are then reclassified by year of application rather than by year of grant. One of our tasks was to be sure that we had all the subsidiaries and names used by a particular corporation. For this purpose we scrutinized the alphabetical index of patenting organizations provided by OTAF and checked it against the list of firms' subsidiaries given in the *Dictionary of Corporate Affiliations* (National Register 1972, 1976) and a list of past mergers given in *Mergers and Acquisitions* (1974–77). If a firm had acquired another firm during this period, we added in the patents of the acquired organization (and its R & D expenditures, when known). In a few cases, where the mergers were large and occurred toward the end of the period, we left the two firms unmerged and instead declared the recent (postmerger) years as missing.

Because the patent data are based on patents *granted* during 1969–78, patents by year applied for cannot really be used before 1968. While only less than 1 percent of all patents granted is granted within the year of application, about 10 percent are granted in the following year. Thus, only about 89 percent of the patents applied for in 1967 would appear among the patents counted by us. Similarly, one probably cannot use the patent data by year of application after 1975, since it takes about four years after the application before more than 96 percent of the patents applied for in that year will be eventually granted are actually granted.[15] Thus, at best, we have about eight or nine years of usable patent data. In most of the analyses we used the eight years, 1968–75. Eight years and 121 firms give us an effective sample size of 968.

Table 3.A.1 gives means and standard deviations for a few of the major variables in the various samples and industries represented in this study. The industrial classification was chosen to approximate the industrial breakdown used by the NSF in its reports. It is clear from this table that these firms are rather large, that the exclusion of firms with R & D budgets of less than half a million dollars makes them even larger, that the size distribution of the firms is quite skewed (standard deviations are

14. Some of the missing years have been interpolated by us. Also, the definition of expenditures reported as R & D by different firms may change over time. Where such changes were obvious or stated in the 10-K forms, we tried to adjust for them. Where we could not and the discrepancies were large, we eliminated the firm from our sample

15. These estimates are based on an unpublished tabulation of patents granted by date applied for, for 1965–77, made available to us by OTAF.

Table 3.A.1 **Characteristics of Sample Firms by Industry: Averages 1963–75 and Standard Deviations**

	Entire Sample			Firms with Complete Data and Min. R & D ≥ 500K		
Variable	*N*	Mean	Standard Deviation	*N*	Mean	Standard Deviation
---------- Ind = 0 ----------						
DEFRND	13	5.678	7.291			
GROPLA72	13	211.994	272.923			
PATS	13	15.788	21.781			
---------- Ind = 28 ----------				---------- Ind = 28 ----------		
DEFRND	21	28.353	32.953	19	31.258	33.362
GROPLA72	21	1053.298	1248.221	19	1162.378	1264.633
PATS	21	92.804	104.502	19	102.520	105.298
---------- Ind = 28.3 ----------				---------- Ind = 28.3 ----------		
DEFRND	20	26.665	20.009	19	28.013	19.603
GROPLA72	20	264.486	206.233	19	277.698	203.002
PATS	20	54.531	40.089	19	57.316	39.151
---------- Ind = 35 ----------				---------- Ind = 35 ----------		
DEFRND	14	17.143	25.603	13	18.407	26.191
GROPLA72	14	327.631	429.287	13	352.300	436.366
PATS	14	47.464	64.881	13	50.990	66.119
---------- Ind = 35.7 ----------				---------- Ind = 35.7 ----------		
DEFRND	13	25.422	30.165	10	32.169	31.521
GROPLA72	13	544.321	767.480	10	665.861	843.293
PATS	13	62.490	98.328	10	79.912	106.895
---------- Ind = 36 ----------				---------- Ind = 36 ----------		
DEFRND	15	15.457	34.068	8	28.427	43.694
GROPLA72	15	393.137	1248.032	8	731.354	1683.749
PATS	15	37.975	68.848	8	70.031	83.442
---------- Ind = 38 ----------				---------- Ind = 38 ----------		
DEFRND	15	25.507	46.550	11	34.452	51.996
GROPLA72	15	352.326	815.241	11	477.772	930.342
PATS	15	63.592	99.897	11	85.920	109.149
---------- Ind = 99 ----------				---------- Ind = 99 ----------		
DEFRND	46	20.489	29.581	41	22.884	30.500
GROPLA72	46	3074.459	10476.016	41	3445.557	11052.898
PATS	46	56.068	90.813	41	61.808	94.687
---------- Combined ----------				---------- Combined ----------		
DEFRND	144	22.612	30.994	121	26.709	32.228
GROPLA72	144	1331.104	6044.733	121	1578.299	6569.299
PATS	144	59.854	84.891	121	70.565	88.661

Note: DEFRND = deflated R & D expenditures, in million dollars.
GROPLA72 = book value of gross plant in 1972, in million dollars.
PATS = number of patents, by year applied for.
Ind = 0 = firms with incomplete data for the whole period.
Ind = 28 = chemicals and allied products, except drugs and medicines.
Ind = 28.3 = drugs and medicines.
Ind = 35 = machinery, except office, computing, and accounting.
Ind = 35.7 = office, computing, and accounting machinery.
Ind = 36 = electronic components and communications.
Ind = 38 = professional and scientific instruments.
Ind = 99 = other manufacturing.

on the order of the means or larger), and that the industrial distribution is quite uneven. The firms represented in the sample are those who reported their R & D expenditures publicly in the 1960s, with drug and chemical firms overrepresented.

The R & D expenditures have been deflated by an R & D "deflator" index constructed along the lines suggested by Jaffe (1972): a weighted average of the index of hourly labor compensation and the implicit deflator in the nonfinancial corporations sector, with .49 and .51 as relative weights.

The main problem with our sample is its peculiar nature. It is based on those companies that reported R & D expenditures in the mid-1960s. Since it is selected on the "independent" variable in this study, one need not anticipate much of a selectivity bias in equations where patents or the market value of the firm are the dependent variables. Also, since much of our analysis will be "within" firms, any fixed selectivity adjustment would be incorporated in the constant term and would not affect our inferences.

Appendix B

The Form of the Patent Equation

Because there was little prior empirical or theoretical research on the R & D-to-patents relation, we began our analysis with an investigation of the functional form of the equation that might connect these two variables in our data.

Functional form questions were examined, allowing the parameters of all estimated equations to differ in each of our seven industries and between firms with large and small R & D departments within each industry.[16] That is, fourteen sets of parameters were estimated. The independent variables included in the estimating equations were a set of time dummies, the current and five consecutive lagged values of both the logarithm of R & D expenditures and R & D expenditures per se, and a set of firm-specific dummy variables (constants). To simplify matters we assumed that the appropriate form of the dependent variable was either $\log(P) = p$ or P itself. Hence log-log, semilog, and linear functions, each with firm and time effects, were all special cases of the model with which we started.

A variant of the Box and Cox (1964) procedure was used to choose the

16. Small firms were defined, quite arbitrarily, as firms whose R & D expenditures over the sample period (1963–1975) fell below half a million dollars in at least one year. The size breakdown had the effect of separating out the recently born science-based firms from the others in the sample and allowed for the possibility that the characteristics of the KPF differed in the firms with smaller, less established, research departments.

form of the dependent variable. It indicated that the logarithm of patents was clearly preferred over the absolute number of patents by the data for each separate grouping and for the sample as a whole. We then asked whether the parameters of the relationship between p and the independent variables within each industry differed between firms with large and small R & D departments. The test statistic was significant at any reasonable level of significance, indicating that the form of the relationship between patents and research expenditures was different for firms with small R & D departments. The twenty-six firms in the small group were dropped from all the subsequent computations reported in this paper. Next, we wanted to know whether the model could be simplified by assuming either that the coefficients of current and all lagged values of R & D, or that the coefficients of the logarithmic forms of these variables, were all zero. The $F^{36,734}$ statistic for the joint significance of the R & D variables in their natural form was a rather small 1.18, whereas that test statistic for the logarithmic form of the R & D variables was a highly significant 3.30. We, therefore, accepted the former hypothesis and rejected the latter and went on to test another simplification: whether or not the seven time dummies could be approximated by a linear time trend. The observed value of the $F^{30,770}$ deviate for this hypothesis was .95, which is below the expected value of that test statistic, given that the time dummies were in fact representing a simple trend. Two other hypotheses were tested but both were clearly rejected by the data. The first was that the distribution of the firm-specific constants was degenerate, that there were no "firm effects." After rejecting this hypothesis we went on to test whether it was reasonable to assume that the firm effects were uncorrelated with research expenditures. It was not. Thus the form of the equation we settled on was rather simple: the logarithm of patents as a function of a time trend, current and five consecutive lagged values of the logarithm of R & D expenditures, and (correlated) firm-specific constant terms.[17]

17. There is one issue which we have not dealt with here because it is not very important in our sample. For observations where $P = 0$, log (P) is undefined. This exposes an underlying truncation problem in our model. That problem, however, is of minor importance for our sample since only 8 percent of the observations are at $P = 0$. This is less than the percentage of observations at $P = 1$ (14 percent), indicating that the truncation problem is not large. It is even smaller for the larger R & D firm sample ($N = 121$) where the zero patents percentage is only three. As a result we treated the whole problem as one of finding a point on the logarithmic scale for $P = 0$. This was accomplished by adding a dummy variable to the independent variables for observations where $P = 0$. The estimated coefficients of this dummy variable are stable across models, implying roughly the value of 0.1–0.7 for the $P = 0$ observations. It does raise the issue, though, of whether our functional form (log-log) is appropriate for low patenting level observations. We intend to investigate explicitly probabilistic models of the patenting process in subsequent work. These issues are discussed in more detail in Hausman, Hall, and Griliches (1984).

References

Bound, J., and B. H. Hall. 1980. The R & D and patents data master file. Mimeo.

Box, G. E. P, and D. R. Cox. 1964. An analysis of transformations. *Journal of the Royal Statistical Society*, series B, 26:211–243.

Comanor, W. S., and F. M. Scherer. 1969. Patent statistics as a measure of technical change. *Journal of Political Economy* 77, no. 3 (May–June):392–98.

Griliches, Z. 1980a. R & D and the productivity slowdown. *American Economic Review* 70, no. 2:343–48.

———. 1980b. Returns to research and development expenditures in the private sector. In *New developments in productivity measurement and analysis*, ed. J. W. Kendrick and B. N. Vaccara, 419–54. Conference on Research in Income and Wealth: Studies in Income and Wealth, vol. 44. Chicago: University of Chicago Press for the National Bureau of Economic Research.

Griliches, Z., and A. Pakes. 1980. The estimation of distributed lags in short panels. NBER Working Paper no. 4, October. Cambridge, Mass: National Bureau of Economic Research.

Hausman, J., B. H. Hall, and Z. Griliches. 1984. Econometric models for count data and with application to the patents–R & D relationship. *Econometrica*, in press.

Jaffe, S. A. 1972. A price index for deflation of academic R & D expenditures. Washington, D.C.: NSF 72–310.

Kuznets, S. 1962. Inventive activity: Problems of definition and measurement. In *The rate and direction of inventive activity: Economic and social factors*, ed. R. R. Nelson, 19–51. Princeton: Princeton University Press for the National Bureau of Economic Research.

Mergers and acquisitions, vols. 8–11. 1974–77. *Journal of Corporate Venture*. McLean, Va.: Information for Industry Inc.

Mundlak, Y. 1978. On the pooling of time series and cross section data. *Econometrica* 46, no. 1:69–86.

Mundlak, Y., and I. Hoch. 1965. Consequences of alternative specifications in estimation of Cobb-Douglas production functions. *Econometrica* 33, no. 4:814–28.

Nadiri, M. I., and G. C. Bitros. 1980. Research and development expenditures and labor productivity at the firm level: a dynamic model. In *New developments in productivity measurement and analysis*, ed. J. W. Kendrick and B. N. Vaccara, 387–412. Conference on Research in Income and Wealth: Studies in Income and Wealth, vol. 44. Chicago: University of Chicago Press for the National Bureau of Economic Research.

National Register Publishing Company. 1972. 1976. *Dictionary of corporate affiliations*. Skokie, Illinois.

Nelson, R. R., ed. 1962. *The rate and direction of inventive activity: Economic and social factors*. Universities–NBER Conference Series no. 13. Princeton: Princeton University Press for the National Bureau of Economic Research.

Pakes, A. 1978. Economic incentives in the production and transmission of knowledge: An empirical analysis. Ph.D. diss., Harvard University.

Rapoport, J. 1971. The anatomy of the product-innovation process: Cost and time. In *Research and innovation in the modern corporation*, ed. E. Mansfield, J. Rapoport, J. Schnee, S. Wagner, and M. Hamburger, 110–35 New York: Norton.

Scherer, F. M. 1965. Firm size, market structure, opportunity, and the output of patented inventions. *American Economic Review* 55, no. 5:1097–1125.

Standard and Poor's Compustat Services, Inc. 1980. *Compustat II*. Englewood, Colorado.

Taylor, C. T., and Z. A. Silberston. 1978. *The economic impact of the patent system: A study of the British experiment*. Cambridge: Cambridge University Press.

Temin, P. 1979. Technology, regulation, and market structure in the modern pharmaceutical industry. *Bell Journal of Economics* 10:429–46.

Wagner, L. U. 1968. Problems in estimating research and development investment and stock. In *Proceedings of the business and economic statistics section*, 189–98. American Statistical Association. Washington, D.C