How special is the special relationship? Using the impact of US R&D spillovers on UK firms as a test of technology sourcing

Rachel Griffith, Rupert Harrison and John Van Reenen

A. Appendix: Data

In order to implement our empirical strategy we need to measure three types of information: the location of firms' innovative activity, firms' productivity performance, and the domestic and foreign spillover pools available to companies. We use data from the US Patent Office (USPTO), firm accounting information, and OECD data on industry level R&D expenditure, to measure each of these respectively.

A.1. Innovative activity

The NBER patent citations data file contains computerised records of over two million patents granted by the USPTO between 1901 and 1999 (available on the NBER web site). We use data on patents applied for after 1975, as information on citations are only available for patents applied for after this date. We combine these data with firm accounting data from the Datastream on-line service, which contains information on sales, employment, investment, capital, R&D and the components of value added.[1]

A.1.1. Inventor location

Patents identify the address (including country) of the inventor(s) listed on the patent application. Table 1 (in the main text) shows the lead inventor's location for the 38,160 patents matched to our sample of 188 UK firms listed on the London Stock Exchange in 1985. We use the share of the firm's patents where the lead investor is located in the US (W_i^{US}) and the number of patents with the lead inventor in the US (P_i^{US}). The average varies across industries, with the highest average shares in Office, Accounting and Computing Machinery (47.5%), Radio, Television and Communication Equipment (47.2%) and Food, Beverages and Tobacco (46.4%). The lowest shares are in Textiles, Leather and Footwear (12.7%), Other Transport Equipment (24.5%) and Basic Metals (28.7%).

A.1.2. Patent Citations

We use data on patent citations to refine our measures of the location of firms' innovative activity. The 38,160 patents matched to our sample of UK firms make

[1] More details of the matching between the datasets can be found in Nick Bloom and John Van Reenen (2002).

275,013 citations to other patents, an average of 7.2 citations made by each patent. Of these 275,013 citations, 236,367 have information on the location of the lead inventor of the cited patent. Because we are interested in whether firms are benefitting from external knowledge that has not been generated within the same firm we exclude self-citations (where a patent cites another patent that is owned by the same firm). 8.5% of all citations in our sample are made to patents owned by the same patenting subsidiary (or "assignee"), while a further 1.4% of all citations are made to a different assignee that is nevertheless part of the same parent firm.

Table A1 shows a cross-tabulation of the location of the citing and cited inventor for the 209,090 non-self citations in our sample. It is important to remember that all of these citations were made by patents that are owned by UK firms, even if the inventor was located in the US. Only 6.9% of citations made *by* UK inventors are made *to* another UK inventor, while 59.9% are made to a US inventor. In contrast, 71.5% of citations made by US inventors are made to other US inventors, while only 3.2% are made to UK inventors. This probably illustrates both the fact that the data is from the US patent office, but also the dominant global position of the US in innovation. This provides preliminary evidence that most patents owned by UK firms, but invented by an inventor located in the US, are building on knowledge created by other inventors located in the US. When we look at self-citations to a patent that is owned by the same parent firm (not shown) the percentages in the diagonals (for example a UK inventor citing another UK inventor) are much higher. We also see that, even within firms, the transfer of knowledge from the UK to the US appears to be small compared to the transfer of knowledge within the US.

A.1.3. Patent Application dates

We also use information on the application dates of each citing and cited patent in order to refine our measures of the location of firms' innovative activity. In particular we look at citations made to patents that were applied for within the last three years. Table A2 shows the same cross-tab of the country of the citing and cited inventor for all non self-citations of this type. The proportions are similar to those in Table A1, although UK inventors are slightly more likely to cite other UK inventors than before, and US inventors are less likely than before to cite other US inventors.

A.2. Firm Accounts data

We use data on firms that are publicly listed firms on the London Stock Exchange and whose primary sales are in manufacturing and who report some R&D between 1990 and 2000. All data relates to the firms' consolidated worldwide accounts. Observations with missing data, firms with less than five consecutive observations over 1990 - 2000, and firms for which there were jumps greater than 150% in any of the key variables (capital, labor, sales) were dropped. Data on value added, labor (DS Item 219) and R&D expenditure (DS Item 119) comes from the Datastream On-Line service (DS). Capital is estimated as a replacement value using the method described in Bond and Meghir (1994). Although these are "UK firms" in the sense that they are listed on the London Stock Exchange, a key feature of the data is that it relates to the firm's global activities. Value added is the sum of total employment costs (DS117), operating profits (DS137), depreciation (DS136) and interest payments (DS153).[2]

The initial sample is all firms listed on the LSE in 1985 with names starting with the letters A-L, plus any of the top 100 UK R&D performers not already included. The sample includes 415 firms, 266 of whom had taken out at least one patent between 1975 and 1998. All these firms' subsidiaries were identified using *Who Owns Whom* by Dun and Bradstreet in 1985.[3] Firms who entered the sample after 1985 were matched based on their date of entry. All the subsidiaries were then matched by name to the USPTO.

In the UK most firms did not report R&D expenditure before 1989, and so the analysis is restricted to the years 1990-2000.[4] An R&D capital stock was constructed using a perpetual inventory method and an assumed 15% rate of obsolescence (Griliches, 1979, and Hall et al, 2004).

Industry codes for UK firms are at the 3-digit level. We matched 3-digit SIC80 codes to 2-digit ISIC Revision 3 codes for the purposes of assigning firms to a 2-digit industry.

After cleaning our data we have a sample with 1794 observations on 188 firms, 141 of which are matched to at least one patent. Table 2 in the main text reports summary statistics. On average, firms in our sample have applied for 240 patents.

To construct the proportion of sales that are made abroad (S_i) we use item 190F from Datastream. We do not have this data for every year, on average we have it for 4 years per firm, mostly during the mid-1990s. Sales are given by region, but the definition of region is left up to the firm to report. We do two things, (i) we take all firms that report sales in "United States", "North America" or "The Americas" (or derivatives of these names) and calculate the share of sales in the US, (ii) use all sales that are in the "UK" to calculate the share of foreign sales.

[2] The first two items dominate this measure.

[3] As with other matches this has the disadvantage that we do not track changes in ownership over time. This is inevitable given the labor intensity of the data matching exercise. Another issue is that we do not track the sales of patents from one firm to another (this may cause us to overestimate the proportion of UK inventors in the US if UK firms buy many US patents). Fortunately such non-M&A related patent sales appear to be a relatively rare event.

[4] Even after 1989 when a firm reports zero R&D it is not clear that this corresponds to a true zero, although it is unlikely to perform a large amount of R&D. In the results presented in this paper, a dummy variable was used to denote reported zero R&D expenditure, but the results are not sensitive to the exact treatment of reported zeros.

A.3. Industry level data - R&D Spillover pool

The domestic and foreign spillover pools were constructed using the OECD's Analytical Business Expenditure on R&D dataset (ANBERD, 2002). This contains information on R&D spending at the 2-digit manufacturing industry (ISIC Revision 3) for all OECD countries. A stock measure was constructed using a perpetual inventory method and an assumed 15% rate of obsolescence,[5] with a starting year of 1987. Although there are various problems with using industry-level measures this data has the crucial advantage for our purposes that it contains R&D expenditures by *geographical location* of the R&D activity. This would be extremely hard to re-create using data on firms' reported R&D as very few firms decompose R&D into a foreign and domestic element. Our measure also has the advantage of including all R&D carried out in each industry in each country, and not just the R&D of the other sampled firms. We also use data on 2-digit industry level value added taken from the OECD's Structural Analysis database (STAN, 2003). Value added price deflators at the two digit level are also from this source. In addition, we use three digit value added from the NBER productivity database and from the UK PACSTAT data (similar findings were uncovered from 3 and 2 digit analysis).

A.4. Technological Proximity Measure

We constructed a measure of technological proximity between our UK firms and US industries following the Jaffe(1986) method. We allocated all R&D performing Compustat firms to a two digit industry and calculated the average technological profile using the average share of patents in each of the 623 technology classes in the USPTO. We then calculated the uncentered correlation coefficient between each of our UK firms and the US industry. The technological proximity formula following Jaffe (1986) between firm i and industry, where firm i is in industry j, is

$$PROX_{ij} = \frac{T_i T_j'}{(T_i T_i')^{\frac{1}{2}} (T_j T_j')^{\frac{1}{2}}},$$

where $T_i = (T_{i1}, T_{i2}, ..., T_{i623})$ is a vector whose elements are the proportion of patents over the 1975 to 1989 period in each of 623 (labelled N-class) technology classes in the USPTO. $PROX_{ij}$ is the uncentered correlation. Compared to the original Jaffe (1986) paper and its descendents we are treating US industry j as a "pseudo" firm. We also tried an alternative measure using all patents among Compustat firms not distinguishing by industry.

[5] We experimented with other depreciation rates but the results were not significantly changed.

B. Appendix: Econometric modelling strategy

In the main text we compare results from two alternative approaches to the problems associated with estimating a production function, a GMM method (Richard Blundell and Stephen Bond, 2000) and the popular "OP" method (Stephen Olley and Ariel Pakes, 1996). These approaches are based on different assumptions and have different strengths and weaknesses[6]. The OP approach has a more flexible form for the "not so fixed" effect of the unobserved heterogeneity (allowing it to evolve over time as a Markov process). The GMM approach allows for a permanent component to unobserved heterogeneity and for the transitory component to be contemporaneously correlated with labor, physical capital and R&D. This Appendix gives some more detail on each method.

B.1. SYS-GMM

Consider a simplified form of the production function

$$y_{it} = \alpha x_{it} + \varepsilon_{it} \tag{B.1}$$

where x_{it} is an endogenous input and the residual productivity term takes the form

$$\varepsilon_{it} = t_t + \eta_i + u_{it}. \tag{B.2}$$

Year dummies (t_t) control for common macro effects, the unobservable firm component (η_i) is allowed to be correlated with the factor inputs (l_{it}, k_{it}, r_{it}), but assumed uncorrelated with the location of innovative activity (W_i^{US}, W_i^{UK}) and all industry level variables, and the residual productivity shock (u_{it}) may be correlated with the factor inputs. Assuming no serial correlation in the u_{it} process yields the following moment conditions

$$E[x_{i,t-s}\Delta u_{it}] = 0 \tag{B.3}$$

for $s \geq 2$.[7] This allows the use of suitably lagged levels of the variables to be used as instruments after the equation has been first differenced. We test for first and second order serial correlation using an LM test, shown at the base of the GMM columns. If there is higher order (but finite) serial correlation in the u_{it} process longer lags can still be used as instruments.

The first differenced GMM estimator has been found to have poor finite sample properties when the endogenous variables are highly persistent, because the lagged

[6] See Zvi Griliches and Jacques Mairesse, 1998, for a discussion and more recently Steven Bond and Måns Söderbom (2005) and Daniel Ackerberg et al (2004,2005)

[7] If there is serial correlation in the error term this can be dealt with by using longer lags as instruments. For example, if $u_{it} \sim MA(1)$ lags dated t$-$3 and earlier will be valid instruments.

instruments are often weakly correlated with the first differences of the endogenous variables. If we are prepared to make assumptions on the initial condition that $\mathrm{E}[\Delta y_{i2}\eta_i] = 0$ and $\mathrm{E}[\Delta x_{it}\eta_i] = 0$ then additional moment conditions become available.[8] The additional moment conditions take the form:

$$E[\Delta x_{i,t-s}(\eta_i + u_{it})] = 0 \tag{B.4}$$

for $s = 1$ when $u_{it} \sim MA(0)$. This means that lagged differences of x can be used as instruments in the levels equations. We test the validity of the additional moment conditions using a Sargan difference test. The levels equations and differenced equations are stacked in a system, each with its appropriate instruments.

We assume that all time varying firm-level variables are endogenous (labor, capital and R&D), whereas all industry-level variables are treated as exogenous. We also examined specifications where the industry-level R&D stocks are treated as endogenous and the results are not significantly affected. The results are robust to lagging the industry-level variables by one period, in which case they can be treated as pre-determined. We instrument firm-level variables in the differenced equation with their levels dated $(t-2)$ to $(t-5)$ inclusive, and in the levels equation by their first-differences dated $(t-1)$, as well as by all time and industry dummies and all exogenous variables. The standard errors we present allows for arbitrary heteroskedasticity and arbitrary serial correlation. We include full sets of time dummies and industry dummies in all regressions.

B.2. Olley Pakes with R&D

Assume that the production function can be written

$$y_{it} = \alpha_0 + \alpha_l l_{it} + \alpha_k k_{it} + \alpha_r r_{it} + \omega_{it} + \upsilon_{it} \tag{B.5}$$

where ω_{it} is the unobserved productivity state which is assumed to evolve as a first order exogenous Markov process and υ_{it} is a serially uncorrelated additional productivity shock. This is equation (**??**) with $\gamma_{i1} = \gamma_{i2} = 0$ and $a_{it} = \omega_{it} + \upsilon_{it}$. The R&D stock ($r_{it}$) and physical capital stock (k_{it}) are quasi-fixed and labor is completely variable. At the beginning of the period t, firm i observes its productivity state ω_{it} and capital stocks. The key difference between ω_{it} and υ_{it} is that ω_{it} is a state variable and affects investment decisions whereas υ_{it} does not.

The firm sets labor and chooses the level of investment in physical capital (I) and R&D. This investment takes one period before it is effective in raising the (deterministic) capital stock and therefore output. Capital stock at time t is therefore determined by decisions made at $t-1$. The additional shock υ_{it} is

[8] Stationarity of y_{it} and x_{it} is sufficient (but not necessary) for these conditions to hold. What is essential is that the first moments of the endogenous variables are time invariant conditional on the time dummies. The higher order moments are unrestricted.

then realized after these choices are made. We ignore selectivity for expositional purposes but introduce this at the end of the section.

The key insight of the OP algorithm is to use the monotonicity of the investment policy function in unobserved productivity. This can be used to obtain consistent estimates of the parameter on labor (α_l) at stage 1 and we then use these at stage 2 to obtain the capital coefficients (α_k, α_r).

B.2.1. Stage One: Estimation of the coefficient of the variable input.

The estimation strategy is to control for the unobserved productivity shock non-parametrically by exploiting the strict monotonicity of the investment policy function. Inverting the investment policy function enables us to write unobserved productivity as:

$$\omega_{it} = \widetilde{\omega}_t(i_{it,}k_{it}, r_{it})$$

Substituting this expression into the production function (B.5) gives

$$y_{it} = \alpha_l l_{it} + \phi_t(i_{it,}k_{it}, r_{it}) + \upsilon_{it} \tag{B.6}$$

where

$$\phi_t \equiv \phi_t(i_{it,}k_{it}, r_{it}) = \alpha_0 + \alpha_k k_{it} + \alpha_r r_{it} + \widetilde{\omega}(i_{it,}k_{it}, r_{it})$$

We do not know the functional form of ϕ_t so we use a series estimator to approximate it.[9] Estimation of equation (B.6) gives a consistent estimate of α_l and estimates of the unknown function ϕ_t.

B.2.2. Stage Two: Estimation of the coefficients on the quasi-fixed inputs.

First, note that we can decompose the productivity term into a part that was expected given the information at $t-1$ (J_{t-1}) and an unexpected component, ξ_{it} :

$$\begin{aligned} \omega_{it} &= E[\omega_{it}|J_{it}] + \xi_{it} \\ &= E[\omega_{it}|\omega_{it-1}] + \xi_{it} \\ &= g(\omega_{it-1}) + \xi_{it} \end{aligned} \tag{B.7}$$

The second line of equation (B.7) follows from the assumption that productivity follows a first order Markov process. By construction ξ_{it} is uncorrelated with J_{t-1} and it can be thought of as the innovation in the ω process between $t-1$ and

[9] Steve Olley and Ariel Pakes (1996) and Jim Levinsohn and Amil Petrin (2003) find that the fully non-parametric estimator of ϕ_t gives similar results to the series estimator. We found that higher order series expansions (instead of our preferred fourth order) made little difference to the results.

t. In the final line of (B.7) we replace the expectation of productivity with a non-parametric function $g(.)$.

Rearranging equation (B.6) after we have an estimate of the coefficient on the variable input (α_l) gives

$$y_{it}^* \equiv y_{it} - \alpha_l l_{it} = \alpha_0 + \alpha_k k_{it} + \alpha_r r_{it} + \omega_{it} + \upsilon_{it}$$

Using equation (B.7) this can be re-written as

$$\begin{aligned} y_{it}^* &= \alpha_0 + \alpha_k k_{it} + \alpha_r r_{it} + g(\omega_{it-1}) + \xi_{it} + \upsilon_{it} \\ &= \alpha_0 + \alpha_k k_{it} + \alpha_r r_{it} + g(\phi_{t-1} - \alpha_0 - \alpha_k k_{it-1} - \alpha_r r_{it}) + \xi_{it} + \upsilon_{it} \end{aligned} \quad \text{(B.8)}$$

Since $\xi_{it}+\upsilon_{it}$ is uncorrelated with k_{it} and r_{it} and we have estimates of ϕ_{t-1} from the first stage, equation (B.8) can be estimated by Non-Linear Least Squares[10].

There are alternative ways to build R&D into the OP model. For example, Buettner (2003) allows past R&D to stochastically effect the future productivity state in addition to ω_{it-1}. We found similar results using Buettner's approach to the method detailed here (see also Daniel Ackerberg et al, 2005, for other suggestions on extending the OP approach).

Since we only observe firms that have chosen to continue operating there may be survivor biases. We follow the same approach suggested by Olley and Pakes (1996) in using a non-parametric expansion of the survival probability to control for selection bias at stage 2 (the firm will continue operations if expected profits exceed a critical cut-off). Because the spillover terms are assumed exogenous they are included as additional exogenous variables in the production function. We include industry and time effects in all regressions. We calculate the standard errors though a bootstrapping procedure with 100 replications allowing for clustering by firm.

References

Ackerberg, Daniel, Lanier Benkard, Steven Berry and Ariel Pakes (2005) "Econometric Tools for Analyzing Market Outcomes", Harvard University mimeo prepared for *Handbook of Econometrics*, James Heckman and Ed Leamer (eds)

Ackerberg, Daniel, Kevin Caves and G. Frazer (2004) "Structural identification of production functions", mimeo UCLA.

Bond, Stephen and Costas Meghir (1994), "Dynamic investment models and the firm's financial policy", *Review of Economic Studies*, 61, 197-222.

[10] Jeffrey Wooldridge (2005) suggests implementing the Olley-Pakes model by using the implied moment conditions and using GMM estimation. This is more efficient than the two-step procedure but it is slightly harder to build in the selectivity terms.

Bond, Steven and Måns Söderbom (2005) “Adjustment costs and the identification of Cobb Douglas production functions”, Institute for Fiscal Studies Working Paper W05/04.

OECD (2003), “Structural Analysis Data Base”, Paris

OECD (2002), “Analytical Business Enterprise Research and Development”, Paris

Vernon, R. (1966) ’International investment and international trade in the product cycle’ *Quarterly Journal of Economics*, LXXX, 190-207.

Wooldridge, Jeffrey (2005) "On estimating firm-level production functions using proxy variables to control for unobservables" mimeo, Michigan State University

Appendix Tables

Table A1: Location of citing and cited inventors: non self-citations

Cited country:	UK	USA	Other	Total
Citing country:				
UK	3,978	34,762	19,332	58.072
	(6.9%)	(59.9%)	(33.3%)	(100%)
USA	3,375	75,249	26,570	105,194
	(3.2%)	(71.5%)	(25.3%)	(100%)
Other	1,463	24,431	19,930	45,824
	(3.2%)	(53.3%)	(43.5%)	(100%)
Total	8,816	134,442	65,832	209,090
	(4.2%)	(64.3%)	(31.5%)	(100%)

Notes: all citations made by patents matched to the 188 UK firms in our sample, excluding self-citations (where the citing and cited patent are matched to the same parent firm). The time period is 1975-1998.

Table A2: Location of citing and cited inventors: non self-citations to patents that have been applied for within the previous three years

Cited country:	UK	USA	Other	Total
Citing country:				
UK	817	5,886	4,549	11,252
	(7.3%)	(52.3%)	(40.4%)	(100%)
USA	459	10,905	4,561	15,925
	(2.9%)	(68.5%)	(28.6%)	(100%)
Other	256	4,242	4,828	9,326
	(2.7%)	(45.5%)	(51.8%)	(100%)
Total	1,532	21,033	13,938	36,503
	(4.2%)	(57.6%)	(38.2%)	(100%)

Notes: all citations made by patents matched to the 188 UK firms in our sample to other patents that have been applied for within the previous three years, excluding self-citations (where the citing and cited patent are matched to the same parent firm). The time period is 1975-1998.

Table A3: Summary statistics for UK patenting firms

	Mean	Median	Standard Deviation	Min	Max
Total patent applications	240	40.5	657	1	5820
UK Location Weight	0.354	0.274	0.363	0	1
UK Location + Citation Weight	0.082	0.017	0.145	0	1
UK Location + Citation Within 3 Years	0.019	0.000	0.054	0	0.5
USA Location Weight	0.462	0.425	0.379	0	1
USA Location + Citation Weight	0.417	0.368	0.349	0	1
USA Location + Citation Within 3 Years	0.162	0.134	0.184	0	1

Notes: 141out of our 188 UK firms matched to at least one patent; location weights are constructed as described in the text.

Table A4 Descriptive Statistics for US firms

	Mean	Median	Standard Deviation
Employees	13,760	3,528	38,640
Real Sales ($1000)	3,196	586.4	10,742
Capital per employee ($)	59,407	34,607	81,630
Real sales per employee ($1000s)	193.736	162.843	128.641
R&D expenditure/value added	0.059	0.029	.198
R&D stock/value added	0.237	0.113	0.567

Notes: All in 1995 prices, 570 firms, 5446 observations, 1990-2000

Table A5: Data underlying Figure 1

Industry	Average annual % Growth in US R&D stock	R&D expenditure /Value added in US in 2000 %	Mean annual labour productivity growth for high W^{US} firms (%)	Mean annual labour productivity growth for low W^{US} firms (%)	Difference in mean annual labour productivity growth rate	Observations in UK sample	Observations in US sample
High US-UK TFP gap industries							
31 Electrical Machinery NEC	6.65	10.1	5.76	4.67	1.08	143	354
24 Chemicals (including pharmaceuticals)	5.23	13.2	5.81	5.73	0.07	191	820
32 Communication equipment	4.13	19.4	5.27	6.16	-0.88	138	725
29 Machinery and equip NEC	3.96	5.8	-0.94	-1.70	0.76	277	659
34 Motor vehicles	3.48	16.1	2.31	4.05	-1.73	63	264
30 Computing machinery	2.39	32.1	2.47	5.18	-2.71	20	323
28 Metal products	1.85	1.9	-2.89	1.03	-3.92	104	268
Low US-UK TFP gap industries							
33 Precision instruments	7.88	31.6	5.11	5.91	-0.80	58	696
20-22 Paper, printing and publishing	6.12	1.6	1.05	0.54	0.50	170	607
27 Basic metals	0.71	1.3	4.28	5.01	-0.72	80	168
25 Rubber and plastics	4.64	3	1.53	-0.95	2.48	72	347
17-19 Textiles and footwear	2.19	0.5	-2.67	2.08	-4.76	174	261
15-16 Food, beverages and tobacco	1.07	1.1	0.87	3.09	-2.21	131	283
35 Other transport equipment	-5.08	18.3	7.10	4.69	2.40	73	109
26 Non-metallic minerals	-4.66	2.3	0.97	0.36	0.61	98	132

Notes: TFP is calculated based on a superlative index. Labour productivity is real value added per worker. US R&D stock is calculated using a perpetual inventory method and a 15% rate of obsolescence.