**Online supplement to:**

# LEARNING BY DOING AND THE LOCUS OF INNOVATIVE CAPABILITY IN BIOTECHNOLOGY RESEARCH

Amit Jain
Assistant Professor
D-ETM and the Department of Strategy and Policy
National University of Singapore
Faculty of Engineering
Block E2, #05-10
7 Engineering Drive 1
Singapore 117574.
Tel: (+65) 6516 1638
Fax: (+65) 6776 0755
E-mail: amit_jain@nus.edu.sg

**January 2013**

**CONTENTS**

**Robustness to Forgetting**

Forgetting implies that recent experience has primacy over older experience as the latter is susceptible to be forgotten (Argote, Beckman and Epple 1990; Benkard 2000; Darr, Argote and Epple 1995; Thompson 2007). Given this, I checked whether the results presented in tables 2 and 4 of the manuscript are robust taking into consideration forgetting. To this end, I recomputed individual experience over 50 different periods of accumulation *Y* ranging from *Y*=0 (no accumulation, complete forgetting) to *Y*=25 (no forgetting) in increments of 0.5 years. Thus, all experience accumulated *prior* to the year *Y* is assumed to have been "forgotten".

**Table A1. Forgetting.** Results remain unchanged when forgetting is taken into consideration

| | (1) | (2) | (3) | (4) |
|---|---|---|---|---|
| | **Forgetting** | **Individual experience x Team experience** | **Team experience x Firm experience** | **Team experience x Firm experience** |
| Time threshold of forgetting (years) | 4.0*** | | | |
| | (0.137) | | | |
| L. Individual experience | 0.490*** | 0.403*** | 0.399*** | 0.400*** |
| | (0.023) | (0.019) | (0.020) | (0.019) |
| L. Team experience | 0.052* | 0.060** | 0.0584** | 0.0458 |
| | (0.027) | (0.028) | (0.028) | (0.028) |
| L. Firm experience | 0.147 | 0.140 | 0.133 | 0.135 |
| | (0.116) | (0.117) | (0.117) | (0.117) |
| Individual Experience x Team Experience | | 0.428* | | |
| | | (0.259) | | |
| Individual Experience x Firm Experience | | | 0.094*** | |
| | | | (0.036) | |
| Team Experience x Firm | | | | 0.448*** |
| | | | | (0.161) |
| L. Team Size | -1.365*** | -1.353*** | -1.355*** | -1.341*** |
| | (0.043) | (0.044) | (0.044) | (0.044) |
| L. Firm size | 0.252** | 0.256** | 0.258** | 0.256** |
| | (0.123) | (0.124) | (0.123) | (0.124) |
| Industry patenting | -1.545*** | -1.559*** | -1.560*** | -1.564*** |
| | (0.148) | (0.149) | (0.148) | (0.149) |
| Technology Scope | 0.010 | 0.010 | 0.010 | 0.009 |
| | (0.010) | (0.010) | (0.010) | (0.010) |
| No. of Claims | 0.001 | 0.001 | 0.001 | 0.001 |
| | (0.001) | (0.001) | (0.001) | (0.001) |
| Technology progress | -0.008 | -0.007 | -0.007 | -0.007 |
| | (0.012) | (0.012) | (0.012) | (0.012) |
| Constant | 3.806*** | 3.825*** | 3.823*** | 3.830*** |
| | (0.034) | (0.034) | (0.034) | (0.034) |
| RSS | 5796 | 5806 | 5804 | 5800 |
| R-square within | 0.520 | 0.538 | 0.539 | 0.538 |

**Standard errors in parentheses. All regressions include firm fixed effects, technology class fixed effects, and year fixed effects. N=5,898 and there are 413 firms. The three interactions terms have been multiplied by 1E-5 to have coefficients in the acceptable range.**

***** p<0.001, ** p<0.01, * p<0.05**

The results of the analysis are presented in table A1. In a first step, I determined the duration of experience retention *Y* that provides the best model fit by iterating over different model specifications varying *Y* from 0 (complete forgetting) to 25 (no forgetting) in steps of 0.5. I determined the significance of the forgetting parameter using grid search (Thompson 2007). The value of *Y* that maximizes model fit is *Y*=4 (model 1) and this threshold has a statistically significant effect ($p<0.001$). Further, it may be observed in model 1 that individual experience (H1) and team experience (H2) result in greater productivity, and that firm experience has no effect. These results corroborate earlier findings.

In models 2-4 of table A1, I assume that the time threshold of individual experience retention is *Y*=4 years, and that prior experience accumulated is forgotten. I check whether the interactive effects of the experience variables have a significant effect on innovative productivity as per hypotheses 4a-4c when forgetting is taken into consideration. Each of the three interactive effects improve productivity ($p<0.001$), indicating that the prior results obtained are robust.

**Robustness to Clumping Effects**

One possible critique of the finding that a learning curve exists in innovation is that clumping spuriously produces the observed learning curve effect. Consider that an innovation project results in the creation of a "clump" of several patents. Patents in such a clump are likely to be closely spaced temporally, to have common scientists in teams, and to use the same technologies. Clumping leads to bias as patents that are part of the project and are applied for earlier on (when experience is lower) are liable to account for a larger portion of fixed costs (in terms of unit labor requirement) as compared to patents applied for later on. In the extreme, it may be argued that patents applied for later in a clump do not require additional development time but solely require additional time to file the patent.

To account for clumping, I identified clumps of patents in a firm based on their temporal proximity, team match, and technology match. Consider any two patents that belong to a firm. Temporal proximity of these two patents is the spacing between the application dates of the two patents. Team match is the proportion of scientists in the first patent that are also part of the team that worked on the second patent. Similarly, technology match is the proportion of technology categories

of the first patent that are also used for the development of the second patent. Patents in a clump will be expected to have high temporal proximity, team match and technology match. It is of note that in identifying clumps, I made use of transitivity: i.e. if patent 'A' matches with patent 'B' and with patent 'C', but patent 'B' does not match with 'C', they will still all three be grouped into a clump due to the match with 'A'.

**Table A2. Clumping of patents.** Clumping effects are eliminated in regressions by averaging out the unit labor requirements of patents grouped together in a clump

| | (5) | (6) | (7) | (8) | (9) |
|---|---|---|---|---|---|
| **Spacing (years) (<=)** | **1** | **1** | **2** | **3** | **5** |
| **Team and technology overlap (>=)** | **100%** | **90%** | **80%** | **70%** | **60%** |
| L. Individual Experience | 0.182*** | 0.181*** | 0.170*** | 0.169*** | 0.153*** |
| | (0.019) | (0.019) | (0.019) | (0.018) | (0.018) |
| L. Team Experience | 0.319*** | 0.320*** | 0.304*** | 0.318*** | 0.309*** |
| | (0.029) | (0.029) | (0.028) | (0.028) | (0.027) |
| L. Firm Experience | -0.020 | -0.026 | -0.042 | -0.055 | -0.066 |
| | (0.102) | (0.102) | (0.099) | (0.095) | (0.092) |
| L. Team Size | -1.373*** | -1.375*** | -1.374*** | -1.419*** | -1.411*** |
| | (0.040) | (0.040) | (0.039) | (0.038) | (0.038) |
| L. Firm Size | 0.292*** | 0.296*** | 0.270** | 0.253** | 0.201** |
| | (0.111) | (0.111) | (0.107) | (0.103) | (0.099) |
| Industry Patenting | -0.143*** | -0.143*** | -0.131*** | -0.110*** | -0.089*** |
| | (0.014) | (0.014) | (0.013) | (0.013) | (0.012) |
| Technology Scope | 0.024*** | 0.025*** | 0.015* | 0.013 | 0.009 |
| | (0.009) | (0.009) | (0.009) | (0.009) | (0.009) |
| No. of Claims | 0.002 | 0.002 | 0.001 | 0.001 | 0.001 |
| | (0.001) | (0.001) | (0.001) | (0.001) | (0.001) |
| Technological Progress | -0.001 | -0.000 | -0.001 | -0.001 | -0.001 |
| | (0.000) | (0.000) | (0.000) | (0.000) | (0.000) |
| Constant | 3.328*** | 3.322*** | 2.940*** | 2.650*** | 2.099*** |
| | (0.033) | (0.033) | (0.033) | (0.033) | (0.033) |
| R-sq within | 0.517 | 0.5179 | 0.5286 | 0.5408 | 0.5461 |

**Standard errors in parentheses. All regressions include firm fixed effects, and technology class and year dummies. N= 6722 for all regressions.**
***** p<0.001, ** p<0.01, * p<0.05**

To eliminate the effect of clumping, I averaged out the ULR of all patents in an identified clump, where the ULR of a patent is the aggregate amount of time spent by scientists in a team (listed on a patent). After this correction, patents applied for early on and part of a clump are allocated the same ULR as those applied for later on, and any spurious learning curve effect is eliminated. However, this correction also eliminates any true learning curve effect that may have been present in the clump. Model 5 of table A2 underestimates clumping as it groups together patents into a clump if they have been applied for within 1 year of each other, and if the teams and technologies used are

identical. As one moves towards model 9, however, one overestimates clumping as it groups two patents into a clump if they are spaced 5 years or less apart and have a technology and scientist match of 60 percent only. Table A2 regressions indicate that the learning curve results are robust to both the underestimation (model 5) and the overestimation (model 7) of clumping effects as individual and team experience consistently have a significant effect on productivity ($p<0.001$). This provides further support for a learning curve in innovation even though to a small extent clumping may make the learning curve phenomenon appear to be stronger than it actually is.

In further robustness tests, I used another specification of clumping in which each patent in a clump is allocated a unit labor requirement equivalent to that of the first patent in the clump. The results of this analysis are qualitatively the same as those presented in table A2.

## APPENDICES

### Appendix A. Matching firm names

Data entry in the USPTO dataset includes significant human error resulting in the misspelling of firm names. For example, Genentech Inc., the leading biotechnology firm has been variously spelled as Genentech, Genentech Inc., Genentech In.c, and Genetech Inc. While seemingly minor, these errors are problematic for determining whether two patents belong to the same firm or not. To address this problem, I used an algorithm developed by Bronwyn H. Hall of the University of California, Berkeley to standardize firm names (Hall 2008). These standardized names take into consideration differences in firm names due to name endings, abbreviations, presence and absence of certain common words (the, for, and…) in firm names, and many other similar problems. Thus, Genentech would be matched with Genentech Inc. as the matching procedure standardizes firm names by eliminating superfluous endings such as "Inc.". These standardized firm names served as the basis with which to do both automated and manual firm name matches across the USPTO, Bioscan and Capital IQ databases, and then within the USPTO database.

### Appendix B. Scientist name matching

Scientist names present problems of misspelling similar to those for firm names. What is more important, however, is that different scientists may have the same name. For example, the name John Smith occurs on 1,936 different patents, and needless to say, it is not one unique John Smith that is responsible for this work. Thus, a mechanism is needed to distinguish between two persons with the same name, as well as to cope with errors in their names. Trajtenberg, Shiff and Melamed (2006) developed one such name matching algorithm, which I replicated and modified in order to match biotechnology scientist names. In this method, scientist names are matched on the basis of whether they sound similar (to compensate for spelling errors) using a procedure called Soundex, on the basis of technology classes, location, firm name, and according to whether the scientists of a given patent cite that of another scientist with a similar name (there are a number of other criteria as well). I found that this procedure produces a number of false matches. To eliminate these false matches, I added an

additional control step: the combined first name and last name of a scientist cannot differ from another similar pair by more than one character. This significantly improved the name matching procedure.

**Appendix C. Identifying technological categories using technology classes in USPTO**

The USPTO defines a technological domain by a primary technological class and a subclass. A patent lists a number of class-subclass pairs, which correspond to technologies in which the patent contributes to the state of art in knowledge. This definition of technological categories as USPTO technology classes, however, is too fine-grained to represent knowledge, as there exist in the classification system over 157,000 unique class-subclass pairs. To resolve this problem, I used the hierarchical property of the USPTO classification system to aggregate technology classes into categories.

The USPTO technology classification system arranges technological classes into a hierarchical structure, which is referred to as the indented structure. The indentation of technology classes is represented by dots (•) in the classification (table A3). Each dot indicates a dependency relationship between one subclass and another subordinate subclass which inherits its parent's properties. To illustrate the hierarchical technology class structure, consider the primary class 435 (Chemistry: Molecular Biology and Microbiology) and the structure of one of its subclasses, notably subclass 173.1, concerning the "treatment of micro-organisms or enzymes with electrical wave energy." Subordinate to the mainline subclass 173.1 (which is not indented), there exists a hierarchical structure of subclasses. The indent symbol "•" provides information on this hierarchy. For instance, 173.2• indicates that it is subordinate to subclass 173.1: it "inherits" all features of class 173.1, and in addition is also "Enzyme treated." Similarly, class 173.6••• is subordinate to class 173.5••, which in turn is subordinate to class 173.4•. Using the indentation property, the system of subclasses can be represented as a hierarchy (figure A4).

In all, there exist in the USPTO 12 levels (0-11) in the hierarchical technology categorization system (table A5). There are 471 primary classes, and 14,541 different level 1 mainline subclasses. At level 11 and above, there exist a total of 157,094 class-subclass pairs. This is too fine grained a classification for the purposes of modeling knowledge accumulation. Since lower level classes inherit

the properties of higher level classes in the classification system, I created technological categories by collapsing all subclasses at the levels 2 to 11 in the hierarchy (table A5) to that of the mainline subclass level 1. This results in a total of 14,541 different technology categories encompassing innovative activity in *all* industries, including areas unrelated to biotechnology such as semiconductors, glass manufacturing and railway rolling stock. Biotechnology innovation takes place in a subset of 1,613 technological categories. Learning by doing and forgetting occur in each of these categories.

**Table A3 The system of indentation present in USPTO technology classes**

| (Sub)Class | Indent | Description |
|---|---|---|
| 435 | | Chemistry: Molecular biology and microbiology |
| 173.1 | | Treatment of micro-organisms or enzymes with electrical or wave energy (e.g., magnetism, sound waves, etc.) |
| 173.2 | • | Enzyme treated |
| 173.3 | • | Modification of viruses (e.g., attenuation, etc.) |
| 173.4 | • | Cell membrane or cell surface is target |
| 173.5 | •• | Membrane permeability increased |
| 173.6 | ••• | Electroporation |
| 173.7 | •• | Lytic effect produced (e.g., growth enhancement or increased production of microbial product) |
| 173.8 | • | Metabolism of micro-organism enhanced (e.g., growth enhancement or increased production of microbial product) |
| 173.9 | • | Concentration, separation, or purification of micro-organisms |

**Figure A4 The hierarchical categorization system of USPTO technological classes**

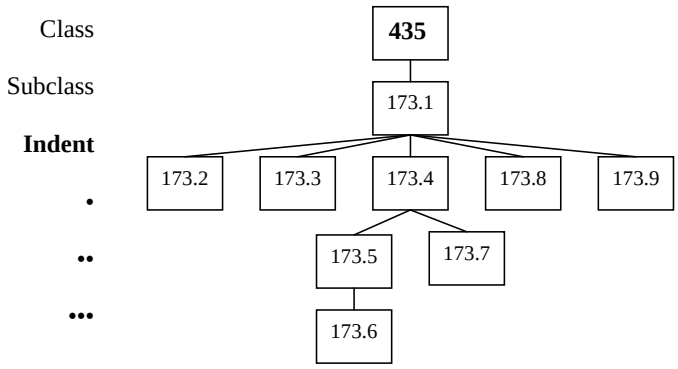


**Table A5 Technology class hierarchy and the number of subclasses[a]**

| Level | No. of distinct subclasses |
|---|---|
| 0 | 471 |
| 1 | 14,541 |
| 2 | 48,164 |
| 3 | 92,042 |
| 4 | 125,099 |
| 5 | 143,296 |
| 6 | 151,444 |
| 7 | 154,706 |
| 8 | 156,161 |
| 9 | 156,767 |
| 10 | 157,000 |
| 11 | 157,094 |

[a]Level 0 = primary class; level 1 = mainline subclass

**References**

Argote, L., S.L. Beckman, D. Epple. 1990. The Persistence and Transfer of Learning in Industrial Settings. *Management Science* **36**(2) 140-154.

Benkard, C.L. 2000. Learning and Forgetting: The Dynamics of Aircraft Production. *American Economic Review* **90**(4) 1034-1054.

Darr, E.D., L. Argote, D. Epple. 1995. The Acquisition, Transfer, and Depreciation of Knowledge in Service Organizations: Productivity in Franchises. *Management Science* **41**(11) 1750-1762.

Hall, B.H. 2008. The Patent Name-Matching Project. Online at: http://elsa.berkeley.edu/~bhhall/pat/namematch.html.

Thompson, P. 2007. How Much Did the Liberty Shipbuilders Forget? *Management Science* **53**(6) 908-918.

Trajtenberg, M., G. Shiff, R. Melamed. 2006. The "Names Game": Harnessing Inventors' Patent Data for Economic Research. NBER Working Paper No. 12479.