Muddles and Models of Management:
Sorting Out the Puzzle of Firm Heterogeneity

Sidney G. Winter*

A rapidly expanding body of high-quality econometric research has underscored the empirical puzzle of firm heterogeneity and produced new insights into the underlying causal factors. This body of research provides a basis, and presents a stimulus, for a careful look at the conceptualization of productive activity – in particular, for examining the differences that separate the basic economic perspective (i.e., production sets and functions) from alternatives derived from management theory, technology studies, and evolutionary economics, and more specifically from the theory of organizational capabilities.. The latter approaches have in common the attribute of being more "situated," i.e., disposed to see production situations as shaped by spatio-temporal, organizational and cultural contexts. The heterogeneity puzzle offers the important challenge of relating these different views in complementary and fruitful ways, and the present paper is an attempt in that direction.

Preliminary Draft, January 2012, not for quotation.

* Deloitte and Touche Professor of Management, Emeritus, The Wharton School, University of Pennsylvania, Philadelphia, PA 19104-6370. Contact: winter@wharton.upenn.edu.

Muddles and Models of Management:
Sorting Out the Puzzle of Firm Heterogeneity

Sidney G. Winter*

"…it is not the business of economists to teach woollen manufacturers how to make and sell wool, or brewers how to make and sell beer, or any other business men how to do their job." – A.C. Pigou (1922)

That the business firms of the world display remarkable diversity is a fact directly accessible to the casual observer. The not-so-casual observer can encounter still more diversity by attending more closely, by exploring far from the main roads of the developed world, or perhaps by spending a considerable period of time "inside" firms. In addition to these possibilities of direct observation, the extent of the diversity is illuminated by a vast range of (largely qualitative) research from the numerous disciplines and research fields that have been concerned with the inner workings of firms to a greater extent than has been traditional in economics – business history, management, technology studies, organization studies, and operations research among others. (Scholars in some of these fields *do* think that "telling business men how to do their job" is part of their own business.) Would anyone who was cognizant of a reasonable fraction of this information on diversity be likely to sum it up by saying, "It looks as though all these firms are about equally effective and must be managed about equally well"? It seems unlikely.

That big fund of information provides, however, little quantitative grip on the phenomena. Without quantitative empiricism, grounded in measurement techniques that are relatively well-defined and widely applicable, it is hard to move beyond the impressionistic level. To say something about the magnitude of the diversity, and then to explore its sources, structure and implications, it is necessary to turn that "diversity" or "heterogeneity" into variance, a task that begins with the question "variance of what?" and goes next to the questions

of measurement technique. The dominant (though certainly not the only) answer to the "what?" question has been total factor productivity (TFP), in some form.

There has been a great burgeoning of research of this kind in recent decades, driven by the strongly complementary influences of advances in econometric theory and information technology (i.e. the computational means to put the theory to work), and most crucially by new access to large micro-level data sets of reasonably high quality – or at least, data sets that are improvable to that level with sufficient care and attention (for overviews see {Bloom, 2010 #662, (Syverson 2011. The current excitement of this field derives not only from the influence of these general drivers, but from innovations in observational techniques that were hardly imaginable a short time ago – including random-assignment experimental interventions in individual firms, and plausible approaches to the direct measurement of the quality of management.

I approach this impressive body of recently-developed evidence as an admirer of the research efforts that produced it and a skeptic about the theoretical commitments that underpin it. Given this pair of reactions, it is reasonable to focus primarily on where they meet, i.e. on the places where the theoretical commitments may be distorting interpretations of the existing evidence or providing guidance for future research that is less helpful than it could be. I will argue that the evidence is not nearly as surprising as it has been said to be, except when seen from the specific viewpoint of those underlying theoretical commitments. Further, the challenging task of understanding the role of management is complicated by the same theoretical commitments that also obscure the phenomena of firm heterogeneity, and for closely related reasons.

Foremost among the referenced commitments are those to the production function as a representation of knowledge, to costless profit maximization as an image of profit-seeking behavior, and to some notion of equilibrium as a way to capture the causal interdependencies among firms. There are many other commitments that are common if not uniform across the econometric research of the subject. Particularly important, given the context, are those that involve direct assumptions of homogeneity in the mechanisms involved typical applications of in the top-level commitments: Not only do firms have production functions, they have *the same* production function, except that (_). Not only do they maximize profit, they do it in the same way, e.g., regarding the same variables as subject to choice, except (_). By making some underlying *homogeneity* fundamental to the theoretical account, such assumptions plainly stack the deck against a prediction of *heterogeneity* – except that (_). Depending on what is listed in the parenthetical "exception conditions" – e.g., production functions differ by Hick-neutral shifts -- the empirical force of the homogeneity assumptions may be greatly attenuated, as more sources of possible of cross-sectional variation are admitted to the picture, allowing a degree of reconciliation with the heterogeneity seen in the data. Whether such a reconciliation is scientifically constructive is deep question, worthy of our attention.

Fortunately, the researchers who produced the evidence are far too capable and far too committed to understanding the phenomena to allow those theoretical commitments to block the door to feasible and otherwise interesting inquiry. (Indeed, the range of hypotheses that have been explored is remarkable; it is striking just how open the door has been.) As a result, the research is generally interesting and informative to someone like me, whose theoretical commitments are entirely different from those espoused as the foundation. This, however, raises the question of how much shaping effect the espoused commitments actually have – or

alternatively, the question of whether those commitments actually have any (refutable) empirical content. Although it might be tempting to argue that the whole structure of the findings can be jacked up and different foundations put underneath, this is not really the case. There are still some potentially resolvable empirical questions on which there is presumably disagreement, at least *ex ante*. Further, the existing commitments seem to steer both the research questions and the interpretations of findings away from some issues that seem interesting and important. Thus, another part of the mission of this paper is to try to sharpen the view of into these areas of disagreement, which inevitably implies a relative neglect of the many questions whose interest and importance are not in dispute. While I do not propose specific hypotheses, I believe many of these relatively neglected questions are not far from the frontier of existing research – and some may, unknown to me, actually be within it.

In the following section, I frame the heterogeneity problem and then relate it to the theoretical issues just alluded to. Section II compares the way the standard production theory treats the knowledge basis of productive activity with the alternative treatment of capabilities theory – a viewpoint considered fruitful in diverse fields in which firm-level account of productive activity form an important part of scholarship, including strategic management, technology studies, business history and evolutionary economics. Section III offers concluding comments.

I. Firm Heterogeneity: Explaining the Surprise

That the accumulating evidence on heterogeneity has been surprising to economists has been firmly attested by (Bloom and Reenen 2007):

> Economists have long speculated on why such astounding differences in productivity performance exist between firms and plants within countries, even within narrowly defined sectors. (P. 1351)

To explain why particular phenomena might be "astounding" to a particular set of observers, it is clearly necessary to consider both the explanation of the phenomenon itself and the sources of the expectations that led the observers to expect something different. I consider both faces of the question here. To evoke the "phenomenon," consider the lead example summarized in Chad Syverson's recent survey article, concluding "… the plant at the 90th percentile of the productivity distribution makes almost *twice* as much output from the *same measured inputs* as the 10th percentile plant." ((Syverson 2011), p. 326, emphasis in original.). Although econometric research has whittled away at the variance by introducing more controls, it is far from the case that it can easily be made to go away.

Regarding the explanation of the phenomenon, it is useful to make a further division of the task into two parts: Where does the heterogeneity come from, and why is it not eliminated by familiar mechanisms? Alternatively, that second clause can be rendered, how can heterogeneity possibly persist at the high measured levels that research discloses?

A. The Sources of Heterogeneity

The sorts of considerations listed in my introductory paragraph seem relevant. The world is a large and diverse place. The observation that there is a lot of cross-sectional variation "out there" is a familiar one, recognized at diverse levels of analysis --from individuals through groups, firms and higher aggregates -- in a variety of more-or-less competitive contexts, and across a range of different attributes and measures. Take the recently completed regular season of the National Football League, for example. The won-lost record of the worst team was 2-14 and of the best is 15-1. The win percentage at about the 90th percentile (rank 4 of 32) was .813 and at the 10th percentile (rank 29) it was .25, for a ratio of 3.25. These teams were competing under conditions of high uniformity of objective (e.g., there is not much strategic incentive to

deliberately risk losing a game, though there is occasionally some), institutional context, and access to the relevant markets. (In my judgment, this is a high degree of uniformity compared to the range of what is found in "narrowly defined sectors" at the national level.) Given these conditions, is the degree of dispersion of results surprising? Is it partly a matter of differences in management quality (from assistant coaches up), or are the managements all clustered close to an objectively defined limit of feasible perfection?[1]

Illustrations of the wide prevalence of high dispersion are extremely abundant, but I will not belabor the point. Perhaps, however, it would be instructive to calibrate the observed dispersions of measured TFP against some of the other examples, taking more care with the quality of the analogies than I have with the NFL example.

What is immediately relevant to my argument is this: It seems clear that the expectations that make the dispersion of TFP values surprising are not framed by reference to the empirical ubiquity of dispersion. They have other sources, and these in general relate to theoretical ideas about how firms and competitive markets are considered to work. Among these, some are more fundamental or firmly held or widely accepted than others. Here is a partial list, amplifying the previous comments on this point.

1) Inputs fall into homogeneous categories and the number of suppliers in each category is substantial enough so that competition among them might plausibly prevail. Further, these categories are fully and, as seen by decision makers, completely labeled: Given a unit of an input, the exact contribution of that unit when introduced in any possible production situation is known – although how that contribution happens may not be known in detail.

[1] The study by (Massey and Thaler 2006) is worthy of consideration here. Incidentally, I acknowledge the possible relevance of a binomial model with p=.5, and also the relevance of the modest sample size of 32, as well as the deep question, "what do you mean by "random"?"

2) Environmental influences on production are invariant across space, or, if varying are fully known and labeled. Spatial variation may or may not be admitted to the model, but in any case does not significantly complicate it. [2]

3) The production set – the set of all feasible production alternatives -- is "given," which means that there is no need to exert costly effort to *invent* "ways of doing things," not even the locally appropriate *details* of a generally familiar "way of doing things," as may practically be required by the spatial and temporal variation of the environment.

4) While producing output is costly, no form of information processing, or decision making or "thinking" is costly when performed by the firm as such– not the effort required to choose input proportions or select suppliers or process the payroll, not the strategic deliberations of the CEO. Because none of this is costly, there are no troublesome opportunity cost issues affecting the allocation of resources capable of these types of information processing.[3]

5) The general advance of science and technology has come to a halt. Hence there is no need to devote effort (costly or not) to surveying the horizon for emerging new alternatives, to evaluating such alternatives, to creating entirely new alternatives, or to assessing the strategic implications of being slow or fast to adopt a new and better way of doing things.

As noted previously, these assumptions stack the modeling deck against the acknowledgment of firm heterogeneity, and in favor of some strong "representative firm" approach to modeling.[4] Further, assumptions 1, 3 and 5 exclude the necessity of whole classes of tasks that real- world

[2] See (Winter 2010) for discussion of the considerations affecting the feasibility of replicating a productive performance across space.
[3] This point rests on the assumption that "economic man is a perfect mathematician," an assumption that is commonly made but rarely made fully explicit (perhaps understandably) by the many theorists who rely on it. But there are important and classic exceptions, notably (Savage 1954), (Marschak and Radner 1972).
[4] By "strong" I mean an approach that allows serious theoretical attention to focus on the means of firm attribute distributions, and addresses the variance by relatively ad hoc measures.

managers clearly perform, while assumption 4 assures that the problem-solving elements of such tasks can be performed costlessly, so long as they are regarded as something done by the firm *per se*, as opposed to an identified input.

As was also suggested in the introduction, one does not expect to see this full suite of assumptions, unqualified, in applied econometric work. In theory textbooks, yes, but in applied work this would be quite self-defeating. What one encounters instead is efforts to accommodate the data by adjusting the *location* in the model, or specification, where the same assumptions, or closely analogous ones, are still in force (as is further discussed below).

The issue of input heterogeneity is an excellent example, especially as it relates to capital equipment. Because the advance of technology denied in 5) is occurring and quite rapid in some cases, and because it may be impossible or difficult to perform the needed quality adjustment on the sorts of data that are typically available (e.g., original acquisition cost), it is clear that studies that do not include such adjustment could be misidentifying an important contribution to productivity growth, treating an unmeasured change in input quality as a TFP improvement. Since it is unlikely that firms are equally disposed to invest in new equipment, an important part of the productivity dispersion story could be missing as well. This issue has been addressed econometrically by methods that, while still very simple in an absolute sense, are a big advance from the "no change" baseline and much different in spirit and implication. Estimates based on the "vintage model" concept, cited by Syverson (2011: 340) suggest very substantial annual rates of quality improvement; as Syverson remarks, "This seems to be an area desperate for further evidence, given its potential importance." (pp. 340-341).

Notice, however, that the retention of the production function framework means that any managerial role, whether in equipment producers or their customers, or the nature of contracts

and communications between them still remains invisible –as does any role for learning-by-doing, the HR practices that might have attracted workers capable of such learning, and so forth. While assumption 5) is gone, and the data better accommodated as a result, presumably assumption 1) remains force, but the list of perfectly-labeled inputs is expanded, likewise the costless identification, assessment and adjustment activities associated with the introduction of the new equipment. The job-destroying implication for management remains.

B. Why Heterogeneity is Often Expected to Disappear

Now take heterogeneity as given. Suppose that, for whatever specific reason, the general diversity of circumstances in the world leads firms to adopt ways of doing things that are significantly different, even when the thing they are doing seems, in output terms, to be pretty much the same – such as producing automobile or personal computers. Possibly some of these differences are attributable to mistaken decisions, in some sense, but defining that "sense" is part of the problem. What is needed here is not a commitment on this error question, or on whether these "ways of doing things" correspond to production functions, but an assumption that the inter-firm differences generate efficiency differences of the kind that would be reflected in differences in measured TFP. Assume also that the firms are in some degree of competition with each other, and the efficiency/TFP differences also are reflected in competitive strength, including profitability.

Imitation.-- Suppose Firm One has a way of doing things that seems to be competitively superior, as measured (at least) by TFP and profitability (measured by rate of return, to put the scale issue aside). What then? Should every firm aspire in principle to adopting Firm One's methods, deferring for the moment the question of whether it is feasible to do so? Clearly, no basis for such a conclusion has been yet supplied, for nothing has been said about the nature of

the space in which all of these firms are going about their business. Perhaps each is making the best of its own idiosyncratic local circumstances -- including price relationships, transportation access, the weather, and other attributes – and Firm One just benefits from having the best local circumstances.[5] For there to be objective "best practices," and thus to clearly justify a general aspiration to adopt Firm One's practices, we need adequate uniformity in the competitive space. So, let us add such an assumption to the developing list, and assume that at least the rank-order of effectiveness (TFP) of the various practices is invariant across the set of sites that firms collectively occupy.

The complex question of the feasibility of imitation is now well posed. It is clear that many aspects of products and processes are observable with minimal effort, and a great deal more can be discovered through more systematic efforts such as reverse engineering, aerial or satellite photography, sustained surveillance and various techniques of industrial espionage. At the advancing frontiers of existing practice, some firms invest in "absorptive capacity" to facilitate their access to various knowledge sources, including the activities of rivals (Cohen and Levinthal 1990). Such systematic efforts are often beyond the means of many small firms, however, and even the "minimal effort" of visiting rivals in the guise of a customer may still be a significant for a sole proprietor who needs to tend her own shop. In short, efforts to learn the practices of other firms are another form of costly managerial activity, and scarcity of managerial resources implies tradeoffs between such activity and all other applications of managerial effort, including improvement efforts that are not informed by the activities of rivals.

Among the most readily observable aspects of a business is its "business model," which can be defined (overly succinctly) as the answers to questions about what is being produced, with

[5] For simplicity, it is helpful to assume that these are single-establishment firms, making "local circumstances" unambiguous. But a multi-establishment firm has a vector of locations and circumstances at a given time; real production has to take place somewhere in real space.

the aid of what input suppliers, and to what customers or classes of customers. Recent literature in management and entrepreneurship has emphasized business model innovation, particularly as a path to profitability (see, e.g., (Baden-Fuller and Morgan 2010) and other papers in that special issue). It is, however, a path that tends to be truncated in time, precisely because of the observability and consequent imitability of business models (Jacobides and Winter 2011). Regardless of the impact on the rewards to the innovator, it is clear that imitation at the business model level is a very powerful force in economic growth, prominently featured in Schumpeter's classic account of how economies grow through innovation (Schumpeter 1934 [1911]). The business model description, however, does not say anything about *how* the result is achieved, in terms of the internal productive practices of the firm. On the average, the "how" information is drastically less observable than the "what," partly because establishments typically have walls and a degree of security, and partly because much of the knowledge is of a quite different kind – as further discussed below. Thus, a business model innovation and a wave of imitation can easily set the stage described by the assumptions set forth above – establishing a heterogeneous collection of rivals doing highly similar things but with idiosyncratic practices, and collectively occupying a competitive a space sufficiently uniform to make the imitability of practices (at the "how" level) a significant issue.

The imitability of business models is one example of a sense or situation in which imitation is a powerful, economically significant homogenizing force. In this example, the force is far from being powerful enough to justify an expectation of low dispersion in TFP. There are many other cases where the force is clearly significant, but the implications for profitability, TFP dispersions and similar considerations are not so transparent. Much confusion has been produced by reckless conclusion-jumping with regard to the role of imitation – proceeding

directly from uncontroversial premises that imitation is going on, firms sometimes work at it, etc., to conclusions about the implications that are far too strong be justified on those premises alone. In general, those conclusions are stated in terms that imply an ultimate outcome of homogenization – with no attempt to express that quantitatively, and of course no reference to the fact that imitation is costly and should be expected to be incomplete for that reason alone. This style of thinking about imitation is probably one of the contributors to the expectation that TFP dispersions should be small.[6]

To introduce a more wide-ranging discussion of imitation, I refer to the following statement by Kenneth Arrow in his classic article of 1962: "…no amount of legal protection can make a thoroughly appropriable commodity of something so intangible as information…" ((Arrow 1962), p. 615). It is easy to imagine that this statement might, ruefully, be regarded as true and even prophetic by Information Age-managers in firms whose products take the form of digitally-recorded information – music, videos, software, and so forth. In contrast to the business model case, technologically-facilitated imitation of such products goes to the value-laden details of their very substance. In principle, process subtleties might forestall imitation, but in the contemporary context this tends to be true in practice only when creators of the information have resisted imitation by some means, including various technological devices and recourse to the legal protections of intellectual property law. Such efforts are costly and, as Arrow indicated, not likely to be fully successful in the long run – or in many cases, the short run.

As in the case of the business models discussion, conceding the centrality of imitation in this particular context does not justify reckless conclusion-jumping regarding the homogenizing effects of imitation. Digitally recorded information is a subset of symbolically rendered

[6] In the strategic management literature, a good example of what I regard as reckless conclusion-jumping about imitation is to be found in a much-cited piece by Michael Porter (Porter 1996). For an antidote, however, see Michael Porter and Jan Rivkin (Porter and Rivkin 1999).

information (including spoken and written language, everyday and technical), which in turn is a subset of the range of information, or knowledge, that is brought to bear in productive activity. In particular, the larger set includes the information stored in human brains that underpins habitual performances of various kinds, ranging from the basic cognitive aspects of perception through language recognition, pattern recognition generally, acquired psycho-motor skills, relational skills and a variety of problem-solving heuristics. Much of this information we now know to be stored in "procedural memory," a neuro-physiological system distinct from the "declarative memory" in which facts reside.[7] There is no question that this is "information" in the economically crucial senses – not diminished through use, and replicable, though at a cost and with some inevitable ambiguity about the precise consequences.[8] (We have lots of touch typists and truck drivers, but they are not all of equal skill.) It is not, however, "information" as economic theorists, or information theorists, typically understand the term. That understanding seems to be confined to the sorts of information that reside in declarative memory, implying it can typically be recorded and communicated in some symbol system, and is or could be recorded in digital form.

This suggests another important source of flawed expectations about the likely extent of heterogeneity in firm practices. The information, or knowledge, that is summarized in production sets is generally conceived in declarative-memory terms, making it difficult to make analytical progress on phenomena involving procedural memory – which all productive performances do. And the bias toward declarative memory is a bias toward the subset of cases to

[7] Richard Nelson and I did not know about procedural memory when we wrote the "skills" chapter of our 1982 book, which precedes the "organizational routines" chapter. (Nelson and Winter 1982) We did, however, have the benefit of Michael Polanyi's discussion of tacit knowledge (Polanyi 1964), which was largely sufficient to the relevant part of our purpose. When the experimental work of Michael Cohen and Paul Bacdayan (Cohen and Bacdayan 1994) brought the psychological research on procedural memory to our attention and showed its relevance to our concerns, it was a revelation … but largely a reassuring one.

[8] See the discussion of "non-standard examples of information economics" in (Winter and Szulanski 2002).

which Arrow's 1962 remark applies – that is to say, a bias toward seeing imitation as relatively easy, and hence as a strong homogenizing force.

When we speak of imitation of (internal) practices, we typically have in mind that the imitator already has its own practices, likely performing the same function as the imitated ones. (It is not like the case of a new entrant imitating an existing business model, going into the same business.) This means that there is a prevailing *status quo*, a short run situation, and the question of the relation of the imitated practice to the *status quo* is an important one. As usual, there is a range of situations. At one extreme, the practice is highly modular relative to the productive system to which it is introduced, meaning the it can be introduced without affecting or requiring adjustment in other aspects. At the other, the new practice it is highly interactive with other aspects of the system, and if the source from which it was drawn is significantly different in those respects, the desired favorable consequences of the change may not be forthcoming – and destructive consequences can occur.

A common variant of this situation occurs when understanding of the practice imitated is limited and does not include recognition of the role of complementary practices present at the imitated site. Thus, while it might have been possible to gain by introducing the new practice together with a piece of its local context, limitations of observation and understanding cause the opportunity to be missed. For example, practices like just-in-time inventory management (JIT), or the famous Toyota andon cord, have the consequence of enhancing the sensitivity of the system to minor disruption, its "tight coupling," which might be thought to be a bad thing. The full package, however, contains multi-tiered responses to the deliberately amplified disruption, aiming at melioration and causal diagnosis in the very short run and at prevention in the

intermediate run. Absent those elements of the package, the amplification of disruption would indeed be a bad thing.

The general point concerning the potential interactions among practices has been noted and explored in a wide range of literatures, and referred to by a variety of terms. It is yet another compelling reason for skepticism about the strength of imitation as a homogenizing force. While there may be widespread awareness of it, queries like "Why doesn't everyone do it that way?" are still often encountered when an apparently superior practice is found in a single firm – suggesting that the presumptive modularity of practices still rules as a default assumption in many minds.

Selection.-- If the firms in our notional competitive environment do not converge in methods by virtue of imitation, perhaps differential growth and the actual exit of weak performers does produce, at least, a substantial shrinking of the share-weighted variance of performance measures – and of TFP in particular. For an evolutionary economist, this restrained (and plausible) conjecture is a distant echo of the much bolder one offered by Milton Friedman many years ago. Friedman famously (at least to some) suggested that the whole apparatus of maximizing behavior favored in mainstream economic theory might be "largely" supported by selection considerations:

> "The process of 'natural selection' thus helps to validate the hypothesis (of maximization of returns-- SW) – or rather, given natural selection, acceptance of the hypothesis can be based largely on the judgment that it summarizes appropriately the conditions for survival." (Friedman 1953), p. 22.

To assess the validity of this conjecture with the aid of some formal modeling was the objective of my own dissertation research long ago ((Winter 1964); for discussion see (Hodgson 1994)). Here, "validity" has to be understood as referring to an asymptotic result in a dynamic system, because that is the sort of thing that modeling can explore. If we ask about what happens while

selection is going on, it is transparently obvious that the result is not to be expected – although some "tendency" in that direction might be seen, to a degree controlled by the specifics. What happens as selection goes on is of more practical relevance, but the asymptotic analysis at least serves to clarify what should *not* be expected to occur via selection as time goes on, i.e., to identify the conceivable tendencies that actually are not predicted by the selection analysis.

The short answer on formal analysis of the Friedman Conjecture is that there is a long list of reasons why its empirical relevance (even asymptotically, if that observation were possible) is very questionable (Nelson and Winter 2002). On the other hand it is certainly possible to confirm it, formally, in an assumed context that is constructed with sufficient care for that purpose (Winter 1971; Winter 1987). I will not recapitulate here a significant fraction of what was discovered in these early investigations and by many subsequent ones, by many authors. For the most part, it suffices to say that a lot of those insights into selection processes are quite relevant to the much narrower question of why the dispersion of observed TFP values is as big as it is, contrary to expectations derived from mainstream theory. Still, there are a few high points that ought to be touched upon.

First, there is the problem of the endogeneity of the environment in which selection occurs. Once we admit the possibility that it is not the case that all possible firms come pre-equipped with optimal responses to all possible environments, it becomes highly relevant that the environment generally changes endogenously as the selection process goes on. Other things equal, output prices tend to fall and prices of sector-specific inputs tend to rise. The performance ranking of practices tends to change as a consequence, and so does the direction of the selection forces. For an illustrative model, suppose that firms make diverse fixed commitments to input proportions between labor and capital, and collectively face an upward sloping labor supply

curve, a constant price of capital, and a downward-sloping demand curve. Capital is fixed in the short run and constant returns to scale prevail in the long; profitable firms grow and unprofitable ones shrink. Measured TFP in any period of course picks up the partial ordering of the input coefficient vectors, but it also picks up the difference in adaptation to the prevailing input price rato.

Second, if "management" or "decision-making" is costly, it matters whether the environment offers benefits commensurate with those costs. It has been well argued in many places that *dealing with change* is the core management problem; a key exemplar is (Penrose 1959). A truly constant environment, of the kind explored in many equilibrium models, does not offer sustained opportunity for the application of managerial skills[9]. No positive-epsilon cost of the ability to deal with different environments is sustained by selection forces in a *constant* environment.

Third, the very large family of evolutionary models of selection processes is, as far as I know, unanimous in its respect (at least implicitly) for the additivity axiom of the axiomatic production theory that is found in advanced textbooks. In the production model of the firm, anything that can be done can be done again and again – absent non-uniformities in the competitive space that, if they exist, ought to be explicitly modeled. In short , (at least) constant returns to scale in production prevail, at least in respect to the scaling of any *specific* way of doing things, in a context that, while possibly changing, is shared by rivals. (There may be non-production costs that do not change in proportion to firm size, such as R&D expenditures.) This set-up is in sharp contrast to mainstream models, such as the one offered by Syverson (Syverson 2011) as an over-arching theoretical structure for his review of the productivity literature. In that model, as in its very similar and much-cited predecessor of thirty years ago (Lippman and

[9] Unless, perhaps ,management is doing process R&D, but that is not the case in the equilibrium models.

Rumelt 1982), it is increasing long-run marginal costs that stand in the way of the selection mechanism, foreclosing the conclusion that (considering selection alone) concentration might increase indefinitely under competitive conditions. Evolutionary models generally produce that conclusion, accept it as a typical asymptotic result under competitive assumptions, and welcome its contribution to understanding the highly-skewed distribution of firm sizes and the impressive empirical power of Gibrat's Law (Geroski 2000).

Fourth, while the institutions of the market economy make exit inevitable for a firm that is performing badly enough, there is nothing inevitable about growth. This is particularly relevant for small firms that operate a single establishment, or a few in a small area, and for which growth would imply a jump in complexity and perhaps a switch to professional management. Thus it is not surprising that some small firms survive and prosper for a long time, but do not expand – and thus do not translate their success into stronger selection pressure on others.

Fifth, the above points direct attention to some stylized facts that theorists ought to take into account, at leastwhen creating models that are intended to inform the interpretation of data. Firm size distributions are highly skewed and the largest firms are much, much larger than the smallest ones, even in narrowly defined sectors. If management is costly and at least some of its important services are relevant firm-wide, then small firms can afford very little management relative to what a large firm can afford (and for the small firms a relatively constant environment will be helpful). Yet, a large fraction of the economic literature, including textbook and advanced theory, and empirical studies, is content with "the firm" as a satisfactory unit of analysis, and seems indifferent to what informational scale economies might imply, give the typical magnitudes of the size differences. Certainly when it comes to profit maximization, all

firms are presumptively on equal footing. That is obviously true in the textbooks. If the econometric research on heterogeneity suggests otherwise, and if the econometricians have learned otherwise it not seem to lead to disavowal of the textbook uniformity assumption.

Another stylized fact, arguably derivable from the previous one, is that life is short on the average for the small firms, though of course there are many excepetions because there are many small firms. It would be "nasty and brutish" as well as short, but for the exit-easing effects of , the bankruptcy laws and limited liability. One of the earliest benefits of the access of academic researchers to the large micro-data sets generated from official statistics was the much clearer picture of the dynamics of the firm-size distribution – specifically, the fact that there is a dramatic level of churn at the small-size and low -age end of the distributions, while relatively tranquility prevails increasingly as attention is directed to higher levels in the distributions of age and size. In spite of this clear evidence, much theoretical discussion and empirical inquiry appears, again, to be structured by the assumption that "a firm is a firm," gives little attention to age and size factors that other empirical research confirms as relevant. Apparently, a firm is a firm in this literature, independent of possibly large size ratios between the largest and smallest firms of the sample. One would think that it would be standard practice to compare separate regressions for the top half and the bottom half of the sample, ranked according to size, to include a report on that largest/smallest ratio in the descriptive statistics, and to approach an assessment of implications by presenting size-weighted results. That would certainly be natural if selection is an important mechanism in the reality of the firms studied.

The foregoing observations suggest, at a minimum, that there is reason to be careful about the empirical computations that relate to the selection effect on aggregate TFP. If the right

high-level right question is about aggregate TFP, the right questions at lower levels relate to input-weighted TFP.

II. Recipe Theory vs. Capability Theory

In general, econometric research on heterogeneity adopts *at some level* and *for some purposes* the standard microeconomic theoretical apparatus of production theory and maximizing behavior. To further explore the consequences of such commitments for understanding heterogeneity and the role of management, I first highlight some key aspects that prevail across the wide range of specific manifestation of the commitments, and then contrast that view and its implications with the alternative of "capability theory" – an alternative that is much more prominent in strategic management and organization studies than it is in economics.

A. Production Theory as Recipe Theory

Production theory can be viewed as a subset of a much larger set of prevalent interpretations of production situations, which I call "recipe theory". Recipe theory is manifested in the world in cookbooks and manuals of great diversity, as well as in the abstract formalizations of production found in microeconomic theory and its applications. The limitations, as well as the strengths, of recipe theory extend across the full range of these manifestations.

In might be argued that to associate production theory with recipes is to pay it an undeserved compliment. Unlike a cookbook or manual, the economists, abstract account of production does not contain representations of the *procedures* that must be followed to produce the output; instead the representation is at the level of input and output quantities – corresponding to the list of ingredients and the yield in a typical cookbook recipe. Most economists do, however, acknowledge that there are procedures in the background, and that the input-output characterization is an instrumental simplification of what full engagement with the

recipe description would suggest. Further, much of what is "wrong" with a cookbook recipe considered as a production theory model can be corrected by a more detailed and economically-informed account of the inputs and outputs, particularly as regards the services of durable equipment, the time cycles involved, and the treatment of by-products and of "disposal". Hence, I think the identification of production theory with recipe theory is justified, and I go forward on the assumption that the steps required to close superficial gaps have been taken; more subtle versions of the gaps come into the discussion below.

What is characteristic of recipe theory in all its particular forms is the claim that bringing together the required inputs – "together" in an appropriate sense, typically geographical – and having the recipe available suffices to make it possible to execute the recipe and make the output appear. This is certainly the implication when a new computer or television or microwave arrives in its box, together with its symbolically-rendered instructions of "set-up" and "getting started." Production theory in economics is even more clearly committed to the basic claim, because production sets and functions are explicitly treated as the full theoretical representation of the required knowledge. As Arrow and Hahn concisely say

> Thus the production possibility set is a description of the state of the firm's knowledge about the possibilities of transforming commodities. (Arrow and Hahn 1971) P. 53.

Firms have production sets, and when they also have the inputs that, per the production set, could be translated into particular outputs, then they can have those outputs if they want them. End of story.

What is missing in the case of the cookbook recipe is the chef and her skills, what is missing for the new computer is the reading comprehension required to understand the instructions in the manual and the skills and manual dexterity required to follow them, and what is missing in economic production theory is an account of the implementation processes -- in

which “management” may play a role. All of these examples illustrate the same thing, a critical limitation of recipe theory. The implementation gaps referred to are all knowledge gaps in large part, and where there is something other than a knowledge gap – e.g., the chef has an impaired sense of smell, the technical competence of the manager is not respected by those who are physically carrying out the activity – those additional features themselves present new questions about what is known, and to whom. In the short, the recipe theory is a failure, insofar as it aspires to isolate the knowledge aspects of the requirements for production in the recipe. (This is clearly the case in production theory, as Arrow and Hahn attested.) There is a lot more to the knowledge requirements than a cookbook, manual or production set can capture.

It is possible, and sometimes useful, to mount a rear-guard action against this line of critical attack by reverting to the question of the input categories and their labeling. Perhaps, for example, the services of chefs and orchestra conductors are available for a price in the market place, likewise those of dishwashers, computer technicians, and CEOs. While this adjustment changes the face and the locus of the knowledge- gap problem in recipe theory, it does not eliminate it. In the case of the theory of the firm, what becomes of the idea that the capacity for optimizing calculation, in fall of those familiar respects somehow resides *in the firm*, which decides what inputs to acquire and what to do with them? The standard assumption of fully informative labeling is drastically implausible for the knowledgeable implementers at all levels, but particularly at the top levels – as the elaborate search processes for appropriate role occupants attest. The difficulties hidden under strong input- labeling assumptions are greatly enhanced by the strong, experience-shaped elements of tacit knowledge in the individual performances, with the attendant inevitable complications of “ambiguity of scope.”

The extent to which inputs get labeled in ways supportive of good decision-making is highly variable across inputs, and strongly shaped by institutional contexts. The "labeling" part often arises incidental to processes that change, or at least categorize, the attributes of the inputs themselves. There is education and training, and then there are degrees, certificates and licenses – the labeling aspect. There are quality checking systems for manufacturing processes that aim to approach within-category homogeneity, and then there are often informative labels on the categories – "factory seconds" for example. As a though experiment, consider the global population of potential productive inputs, just as they are, and imagine that they are all stripped of every institutionally-based label. As an employer of inputs,, you can take into anything that you can independently verify, but you have no recourse to institutionally based back-up -- no records of educational attainment, or verifiable records of claimed experience, no certificates and licenses, including driver's licenses, no exchange-based grading systems for commodities, and so forth. What happens to productive efficiency under those conditions? Yet the conditions of production are unchanged in a physical sense, if not in the production theory sense (given the assumed input categories at its foundation). The conditions of the thought experiment are obviously extreme but the variation in the labeling institutions, internationally and otherwise, is very large, and might well have implications for observed heterogeneity.

The commitments associated with recipe theory encumber understanding of knowledge-related implantation gaps, which link in turn to the imperfections of input labeling, and to the neglect of management's role. These are all related to the basic question "Where does the knowledge reside?, which was among the key challenges to orthodox production theory that Nelson and I identified thirty years ago. Those questions are treated very differently in capability theory, to which I now turn.

B. The Alternative Theory: Organizational Capabilities

In the economics literature, Richardson (1972) may have been the first to employ the term "organizational capabilities" it in relation to the study of firm behavior, and very much in its present sense. In an essay that anticipated later work on inter-firm alliances, networks and supply chains, as well as capabilities, he pointed out a key limitation of the production function construct used in standard economics: "It abstracts totally from the roles of organisation, knowledge, experience and skills, and thereby makes it the more difficult to bring these back into the theoretical foreground in the way needed to construct a theory of industrial organisation." (Richardson 1972) (p. 888) He went on to label the effect of knowledge, experience, and skills with the term "capabilities", and then developed the implications of "…the fact that … organisations will tend to specialise in activities for which their capabilities offer some comparative advantage." (loc. cit.). The unifying thread in the long history of "capabilities" as a term is the emphasis on *what an organization can actually do*, and the importance of the distinction between that question and concepts such as "intentions," "incentives," "motivations," and variations of "having the recipe." As Richardson suggested, economics has long manifested a weak grip on the distinction between capability and these other concepts.

The capabilities approach is very different from standard production theory. It offers a highly "embedded" conceptual account, which is to say that first of all that capabilities are conceived as being created and modified in the course of historical time. More broadly, no aspect of historical and institutional context is necessarily excluded a prior as a possible shaper of firm capabilities. For this reason, the concepts match up easily with the accounts of particular capabilities of particular firms, found in histories and case studies – which, no doubt, is why the approach tends to appeal to historians (see, e.g. (Chandler 1992; Usselman 1993)). That same acceptance of richness and context-dependence naturally tends to create a barrier to the creation of simple analytical parables, but that is not to say that this is

impossible. As in standard theory, there are always choices to be made about levels of details and complexity, depending on what the research question is.[10]

For economists brought up on standard production theory, the closest thing to a familiar path to capabilities is via the short run – long run distinction. The "short run" concept in economics is typically explained in terms of commitments to fixed inputs, and fixed inputs are most typically conceived in terms of durable equipment – although human resources sometimes make the list as fixed or "quasi-fixed" factors of production. Commitments to fixed inputs are certainly part of the capabilities story. Where the difference arises is in the treatment of production techniques and other aspects of firm behavior that standard theory treats as objects of maximizing choice. In capabilities theory (as in evolutionary economics more generally), ways of doing things are part of the capability, often closely related to the traditional fixed inputs (machine operators and their methods along with the machines); they are often governed by organizational routines and, of course, by the exercise of individual skill – not by the sort of direct, profit-oriented, firm- level, active deliberation or calculation that is explained in the economics textbook. So, in capabilities theory, the answer to the question "Where does the knowledge reside?" is that it resides in a lot of places, including individual skills and symbolically-rendered recipes and manuals, but in particular it resides in organizational routines, which are the product of organizational learning. It suffices here to explain "routines" as multi-person skills; see (Cohen and Bacdayan 1994).

This is not the place for a full recapitulation of the whole conceptual system, but it does seem appropriate at this point to offer at least a definition of an organizational capability:

> An organizational capability is a high-level routine (or collection of routines) that, together with its implementing input flows, confers upon an organization's management a set of decision options for producing significant outputs of a particular type. (Winter 2000), p. 983.

In that definition, the reference to "implementing input flows" is intended to evoke, for economists, the usual content and concerns of production theory. And the role of the routines,

[10] For example, (Winter 1971) presents a simple analytical parable about the achievement of a conventional equilibrium outcome in a "capabilities" world.

the building blocks of capability, is to provide the organizational "nervous system" – the learning-based competence that fills the implementation gaps that production theory leaves unaddressed. For further conceptual discussion and many case examples, see (Dosi, Nelson and Winter 2000)

To view firm behavior in this way is not to consider it rigid, unchanging, or uninfluenced by profit-oriented motivation. All three of these issues do appear in a different light when viewed in the skills/routines/capabilities framework, and all pose interesting questions. Reference to the individual skill level should, however, suffice to illustrate what is *not* entailed by capabilities theory at the organizational level. Today you have a limited portfolio of language competence, covering, presumably, quite a small fraction of the world's languages. Does that mean that you are (1) inflexible in the use of the skills you do have, (2) unable to learn new languages, or (3) definitely uninterested in communicating with persons from countries whose language you do not know? No, no and no. What is true, however, is that the heritage of past learning investments in the (tacit) skills of language use is quite durable, and would likely provide some guidance to an aspiring predictor of your (language) behavior for a long time to come. You cannot change your language portfolio overnight by reading a book, although you can carry a phrase book along to compensate partially for the failure to make the necessary investments in the past, or you might hire an interpreter. And so it is, with suitable adjustments, for the behavioral patterns that emerge from organizational capabilities.

I turn briefly to the question of long-run change: There is no "long run" in capabilities theory that it is at all analogous to the long run in the standard theory of production. There is no time when the firm has all the knowledge it will ever need (in its production set) but no durable commitments to any inputs, and all the options are thus open for the maximizing choice. A potential entrepreneur, thinking of founding a business, may indeed have quite a blank slate – but a blank slate corresponds to a largely blank organizational mind, since the knowledge that matters largely comes (for any substantial enterprise) embedded in the inputs or is created through by their shared experience. In a substantial going concern,

the heritage of the past is a much more powerful determinant of both constraint and opportunity – just as in the language skills case. On the other hand, organizations do encounter a flow of variously enticing opportunities for change. Among the most important ones, from a long-run system viewpoint, are the changing technological opportunities that arise, from institutions outside the for-profit business sector, and then are extended and commercialized by businesses. A further advantage of the capabilities approach is that it relates naturally to this important portion of the total knowledge picture (Usselman 1993; Mowery and Nelson 1999).

As remarked by (Bloom, Genakos, Sadun and Reenen 2011) "…management resembles a technology and there can be technical progress in management, just as there is for machines," p. 24. This observation is entirely compatible with the capabilities viewpoint. Management practices settle questions of detail in firm operations just as equipment characteristics, individual skills and organizational routines settle details – and often these are all highly entangled in the generation of firm behavior. If, however, management is also the locus of a firm-wide optimizing viewpoint, as supposed in standard theory, it seems that there is considerable tension between the two conceptualizations. The practices, or routines, typically have limited scope, are only rarely derived from any explicit optimization or cost/benefit calculation. They are, however, certainly real and partly observable.[11] On the other hand, the grand firm-wide optimizing is it not observable; it is an "as if" parable, is it not? Does it contribute anything to the understanding of practices? Or for that matter, to the understanding of performance differences among heterogeneous firms?

[11] I take "routines" to be a subset of "practices." I think of the reliance on procedural memory as defining of routines, and not all practices have that character.

III. Concluding Comments

It is clear that firms do not produce in the theoretical "long run" when the blank, commitment-free slate is in front of them. They produce in the short run with the aid of inputs that bring to the situation at least a part of the knowledge that informs the production process, and in most cases with the aid of a firm history that already provides partial answers to the recurrent questions about of how the organization should go about producing things and addressing the market. The data that underpin the recent studies of firm heterogeneity comes from such sources. Thus, it is clear that the wealth of recent findings on heterogeneity *could be interpreted* as observations on a world described by capabilities theory, though that would leave some of the actual interpretive apparatus un-motivated or perhaps contradicted. If there is a significant theoretical stake in the production-theoretic foundations provided for TFP measurement – it would be good to know what that stake is, and to see whether the solidity of the foundation can be empirical assessed. A more ambitious goal would be identify econometric tests that might discriminate between the standard production-theoretic interpretation and a capabilities interpretation, and that could actually be implemented using kinds of micro-data on which recent productivity research is based.

In my view, it is overly pessimistic to say that no such tests could be derived. But, assuming that nothing persuasive emerges on that front, the "good news" is that it then is costless to accommodate the abundant non-econometric evidence on firm production behavior by adopting capabilities theory as the primary interpretive viewpoint. Among the advantages of that move is that it makes the study of management practices, a natural move – which, in my view, it is not the case in the standard production-theoretic framework.

References

Arrow, K. J. (1962). Economic welfare and the allocation of resources for invention. The rate and direction of inventive activity. R. Nelson. Princton, N.J., Princeton University Press**:** 609-625.

Arrow, K. J. and F. H. Hahn (1971). General Competitive Analysis. San Francisco, Holden-Day.

Baden-Fuller, C. and M. Morgan (2010). "Business models as models." Long Range Planning **42**(2-3): 156-171.

Bloom, N. and J. V. Reenen (2007). "Measuring and explaining management practices across firms and countries." Quarterly Journal of Economics **122**: 1351-1408.

Bloom, N., C. Genakos, R. Sadun and J. V. Reenen (2011). Management practices across firms and countries. Working Paper, Stanford University.

Chandler, A. (1992). "Organizational capabilities and the economic history of the industrial enterprise." Journal of Economic Perspectives **6**: 79-100.

Cohen, M. and P. Bacdayan (1994). "Organizational routines are stored as procedural memory." Organization Science **5**: 554-568.

Cohen, W. M. and D. Levinthal (1990). "Absorptive capacity; a new perspective on learning and innovation." Administrative Science Quarterly **35**: 128-152.

Dosi, G., R. R. Nelson and S. G. Winter (2000). The Nature and Dynamics of Organizational Capabilities. Oxford, Oxford University Press.

Friedman, M. (1953). The methodology of positive economics. The Methodology of Positive Economics. Chicago, University of Chicago Press.

Geroski, P. (2000). The growth of firms in theory and practice. Competence, Governance and Entrepreneurship. N. Foss and V. Mahnke. Oxford, Oxford University Press**:** 168-186.

Hodgson, G. (1994). "Optimization and evolution: Winter's critique of Friedman revisited." Cambridge Journal of Economics **18**: 413-430.

Jacobides, M. G. and S. G. Winter (2011). "Capabilities: Structure, agency and evolution." Organization Science **Articles in Advance**: 1-16.

Lippman, S. and R. Rumelt (1982). "Uncertain imitability: an analysis of interfirm differences in efficiency under competition." Bell Journal of Economics **13**: 418-438.

Marschak, J. and R. Radner (1972). Economic Theory of Teams. New Haven, CT, Yale University Press.

Massey, C. and R. H. Thaler (2006). The Loser's Curse: Overconfidence vs. Market Efficiency in the NFL Draft. New Haven, CT, Working Paper, Yale School of Management.

Mowery, D. C. and R. R. Nelson, Eds. (1999). Sources of Industrial Leadership: Studies of Seven Industries. New York, Cambridge University Press.

Nelson, R. R. and S. G. Winter (1982). An Evolutionary Theory of Economic Change. Cambridge, MA, Harvard University Press.

Nelson, R. R. and S. G. Winter (2002). "Evolutionary theorizing in economics." Journal of Economic Perspectives **16**: 23-46.

Penrose, E. (1959). The theory of the growth of the firm. New York, John Wiley & Sons.

Polanyi, M. (1964). Personal Knowledge: Towards a Post-Critical Philosophy. New York, Harper & Row.

Porter, M. and J. Rivkin (1999). Matching Dell. HBS Case 799-158. Boston, Harvard Business School.

Porter, M. E. (1996). What is strategy? Harvard Business Review. **74:** 61-78.

Richardson, G. B. (1972). "The organisation of industry." Economic Journal **82**: 883-896.

Savage, L. J. (1954). The Foundations of Statistics. New York, Wiley.

Schumpeter, J. (1934 [1911]). The Theory of Economic Development. Cambridge, Harvard University Press.

Syverson, C. (2011). "What Determines Productivity?" Journal of Economic Literature **49**(2): 326-365.

Usselman, S. W. (1993). "IBM and its imitators: organizational capabilities and the emergence of the international computer industry." Business and Economic History **22**(Winter 1993): 1-35.

Winter, S. G. (1964). "Economic 'natural selection' and the theory of the firm." Yale Economic Essays **4**(Spring 1964): 225-272.

Winter, S. G. (1971). "Satisficing, selection and the innovating remnant." Quarterly Journal of Economics **85**(May 1971): 237-261.

Winter, S. G. (1987). Competition and selection. The New Palgrave: A Dictionary of Economics, v. 1. J. Eatwell, M. Milgate and P. Newman. New York, Stockton Press. **I:** 545-548.

Winter, S. G. (2000). "The satisficing principle in capability learning." Strategic Management Journal **21**(Oct-Nov (special issue)): 981-996.

Winter, S. G. (2010). The replication perspective on productive knowlege. Dynamics of Knowledge, Corporate Systems and Innovation. H. Itami, K. Kunisoki, T. Numagami and A. Takeishi. Berlin, Springer-Verlag.

Winter, S. G. and G. Szulanski (2002). Replication of organizational routines: conceptualizing the exploitation of knowledge assets. The Strategic Management of Intellectual Capital and Organizational Knowledge. C. W. Choo and N. Bontis. New York, Oxford University Press**:** 207-221.

References 2