A GENERAL INTER-INDUSTRY RELATEDNESS INDEX

by

David J. Bryce *
Brigham Young University

and

Sidney G. Winter *
University of Pennsylvania

CES 06-31 December, 2006

Abstract

Firm growth and expansion is widely believed to be guided by the desire to leverage existing resources. But which resources? The answer depends largely on context—the peculiarities of industries, firms, technologies, production, customers, and a host of other dimensions. This fact makes pointing to any particular set of resources as the source of expansion decisions potentially problematic and makes more difficult tests of theories such as the resource-based view of the firm. This paper tackles the problem by developing a general inter-industry relatedness index that can be usefully applied across industry and firm contexts. The index harnesses the relatedness information embedded in the multi-product organization and diversification decisions of every firm in the US manufacturing economy. The index is general in that it implicitly varies the underlying resources upon which expansion proceeds with the industries in question and provides a percentile relatedness rank for every possible pair of four-digit SIC manufacturing industries. The general index is tested for predictive validity and found to perform as expected. Applications of the index in strategy research are suggested.

* Thanks go to Connie Helfat, Steven Klepper, Anne Marie Knott, Harbir Singh, Louis Thomas, Jeff Dyer and Robert Hoskisson, as well as participants at the BYU-Utah Winter Strategy Conference, for helpful comments and suggestions. This research was conducted while the first author was a Census Bureau research associate at the Center for Economic Studies. Research results and conclusions expressed are those of the authors and do not necessarily indicate concurrence by the Census Bureau, or by any of the colleagues thanked above. Results appearing herein have been screened to insure that no confidential data are revealed.

> If there are profitable opportunities for increased production anywhere in the economy they will provide for some firm an external inducement to expand. But this alone tells us nothing about their significance for any given firm. [Opportunities] are external inducements to expand only for what might be termed ‘qualified’ firms—firms whose internal resources are of a kind either to give them a special advantage in the ‘profitable’ areas or a least not to impose serious obstacles (Penrose, 1959/1995:86).

1. INTRODUCTION

Profitable expansion opportunities are not uniformly available to all firms. Whether a particular opportunity has profit potential for a particular firm is determined in part by the firm’s stock of resources, either already present or assumable, by which it takes advantage of the opportunity (Penrose, 1959). Thus, the question of predicting which opportunities a firm pursues in expansion is answered by first addressing the question of which resources are to be leveraged. If a theoretical commitment can be made to a class of resources upon which expansion decisions are expected to be carried out, then the operational problem of “which resources” is reasonably straightforward. But if the specific resources relevant to expansion are unknown, difficult to identify, or are expected to change with the opportunity, firm, or industry context, then the problem of resource identification is considerably more complicated. This latter situation arises frequently in broad-based studies of diversification where heterogeneous industries are present in the portfolios of firms under study. Strategy scholars examining the resource based view of the firm (Wernerfelt, 1984; Barney, 1991) have often bounded analysis within homogeneous populations, such as technology firms (e.g. Silverman, 1999), in order to circumvent the complication.

In this paper we propose a remedy by developing a general relatedness index that can be usefully employed to examine firm expansion decisions across industry and firm contexts. The index is general in that it implicitly varies the underlying resources upon which expansion proceeds with the industries in question and provides a percentile relatedness rank for every possible pair of manufacturing industries. Our index harnesses the relatedness information embedded in the multi-product organization and diversification decisions of every firm in the US manufacturing economy for the specific time period on which it is based.[1] In contrast, the most common traditional measures of relatedness or diversification, such as the concentric index or the entropy measure, infer relatedness from the hierarchical structure of the Standard Industrial Classification (SIC) system. Our index uses the SIC system only to categorize industrial activity at the most micro level, which we take to be the 4-digit level. The methodology could be applied to any system that provides an exhaustive classification of activity at whatever is considered to be the micro level, and to any time period for which the requisite data are available.

Patterns of corporate diversification are, we assume, shaped in a fundamental way by a logic of economic efficiency. Opportunities for profitable diversification moves arise because there is some overlap between the resources and capabilities that support the existing portfolio of activities and those that are required in some new line of activity. Such overlaps produce “economies of scope” – a term that we use in a broad sense to cover any and all sources of economic gains arising from the combination of disparate activities in a single firm. We generally presume that such “scope economies” arise in the most fundamental and durable sense from non-rivalrous information that is valuable in two or more different activities. Unlike an amount of underutilized productive capacity of a particular type, or a relationship with a particular distributor whose capacity is limited, underutilized knowledge is leverageable to an indefinite extent. There is no limit to its application that is intrinsic to its own nature, though of course the environment may impose such limits. Our methods do not, however, rely specifically on the assumption that economies of scope have this knowledge-based character.

At any given time, the patterns of corporate participation in different industries reflect the cumulative effect of the operation of this efficiency logic in the past – along, of course, with whatever other causal determinants and random effects may be involved. On this basis, our methodology may be viewed as relying upon the *survivor principle* in that it presumes that what firms actually do makes economic sense. Thus, if a firm is observed to be participating in both industry A and industry B, the observation supports the inference that A and B are “related.” It makes some kind of economic sense for the firm to be doing that (Teece, et al. 1994), and the economy-wide implication of such firm-level sense

[1] Every firm down to a size of three employees.

is what our index seeks to capture. As originally stated by Stigler (1968:73) the survivor principle is that "the competition of different sizes of firms sifts out the more efficient enterprises."[2] In relying on this principle, we do not presume that it operates with great promptness or precision. Rather, we presume that the economic forces shaping the observed reality are diverse both qualitatively and quantitatively. Other causal forces, random effects and organizational inertia may certainly shape the observations when the economic forces when are weak – but this is not so likely, we assume, when they are strong.

Since the starting point of our approach is instances of firm participation in two industries, this work is plainly related to the much-discussed question of firm boundaries or "the nature of the firm" (Coase 1937). We assume that the observation that a firm engages in activities A and B does not merely suggest the existence of affirmative economic reasons for this combination (i.e., relatedness), but also that standing objections to such combinations were overcome in this case. Regarding the specific nature of those "standing objections," we do not make, and do not require, any specific commitment. Certainly the literature of transaction cost economics offers valuable insights on this matter (e.g., Coase 1937; Williamson 1985). Certainly we agree that the fundamental question that Coase derived from Lenin – "Why is the economy not run as one big factory?" (Coase 1991) – must have an economic answer. We do suspect, however, that the historical paths of capability development in firms may have more to do with that answer than transaction cost theorizing seems to allow. In any case, we conjecture that the *absence* of any instance of a firm that does both C and D also makes some kind of economic sense; the question, again, is how controlling the durable economic forces actually are.

The predictive value of our index rests on the premise that the methodology captures fundamental aspects of relatedness among industries, so that the relatedness score it generates is accounted for by relatively durable considerations. In reference to the time period from which it is inferred, it is of course tautological to observe that participation patterns reflect "relatedness" as measured by the index. But in reference to subsequent time periods, the durable features of knowledge structure reflected by the relatedness score remain. If we are correct that the index captures such features, it can be used both as a

[2]

reference standard for relatedness content between industries and as a predictive tool in those settings. (Needless to say, there is no example of quantitative prediction in the domain of science that does not rely on an assumption that something-or-other measured at one time is still holding that same value at a later time.)

To test the predictive validity of the measure, we employ a more conservative test than is represented by examining the direction of corporate diversification directly. Using knowledge-based theorizing, we argue that our index should predict the *mode* of entry of an expanding firm. The test is a demanding one in that it requires the index to distinguish between acquisition and organic expansion rather than simply showing that there is high relative relatedness between activities in the firm's portfolio and the industries the firm actually enters.

The index is applicable to a wide range of problems in strategic management, corporate finance, and economics since it provides a plausible measure, grounded in an economic efficiency logic,of the relative strength of associations between every pair of industries.. Applications of the measure to the study of longitudinal patterns of diversification and firm growth are especially promising because the measure allows consideration of new industries one industry at a time. Intra-portfolio distances are easy to compare and the relatedness distance of an activity outside of the firm is readily assessable. The index lends itself particularly well to examining incremental entry or investment decisions by firms in the context of activities already in the portfolio. The index is not, however, a measure of diversification and so cannot be directly compared to extant diversification measures. The familiar diversification measures are portfolio-level constructs that typically include a crude relatedness component; relatedness measures are used to characterize a more fine-grained relationship between classes of activities.

The "coherence" methodology of Teece et al. (1994) forms the effective starting point for development of the general index. That methodology, however, was not intended to produce a general inter-industry relatedness measure and it therefore has limitations that must be circumvented when it is applied for this purpose. That approach does not, for example, consider the relative importance of activities in a corporate portfolio, instead treating all activities as significant if they appear in the portfolio at all. This is an issue particularly for broadly diversified firms with widely differing participation levels in a variety of industries. The Teece *et al.* methodology also poorly distinguishes the level of relatedness between industries that are not combined in any firm. These and other issues are resolved in this paper.

The discussion proceeds as follows. Section two reviews the literature and highlights differences between measures of relatedness and diversification. Section three outlines some methodological hurdles in developing a general index and proposes solutions. Section four develops the index. Section five offers the test of predictive validity and Section six concludes.

2. DIVERSIFICATION AND RELATEDNESS

The concept of relatedness in strategy research was first employed to assess the linkage between diversification strategy and performance proposed by Chandler (1962). Building on Chandler, developments in strategic management have emphasized that firm portfolios in which businesses are interrelated should produce higher levels of performance than portfolios in which businesses are unrelated (Rumelt, 1974; Montgomery, 1979; Rumelt, 1982; Teece, 1980, 1982; Ramanujam and Varadarajan, 1989). The hypothesis is that combinations of related activities are expected to produce economies of scope in production (Teece, 1982; Panzar and Willig, 1981). These economies are an important potential source of performance differences between firms that pursue strategies of related diversification versus unrelated diversification. Since diversification strategy is an aggregate construct, however, relatedness is typically assessed at the aggregate portfolio level, with differing levels of inter-activity relatedness within the company being combined through some explicit or implicit weighting scheme. Accordingly, the most commonly used measures of diversification contain at least two components: (1) A component that assesses the degree of relatedness among activities; and (2) a component that weights activities, providing greater weight to activities accounting for a relatively greater proportion of the business.[3]

[3] Some argue that the number of businesses should also be included (e.g. Gort, 1962). Lubatkin, Merchant, and Srinisvasan (1993), for example, showed that product count measures of diversification correlate strongly to the

A number of methods have been developed to assess diversification strategies along these lines, beginning with the Wrigley-Rumelt categorizations of diversification type (Wrigley, 1970; Rumelt, 1974; Montgomery, 1979). Categorical measures consider a related-diversified firm to be a firm whose largest single group of related businesses (as assessed by researcher judgment) accounts for seventy percent or more of revenues. Measures that have a long history of use in industrial organization economics (to measure the concentration of sales among firms in an industry) have been adapted to assess the concentration of a single firm's activities among industrial categories. The most familiar measure of this type is the Herfindahl index; which was employed in the diversification context by Berry (1971). Such a measure addresses the relatedness of a firm's activities only in the limited sense that for a given system of categories, within-category relatedness of activities is assumed to be perfect, while between-category relatedness is always zero. Gollop and Monihan (1991) proposed a "generalized index of diversification" that involved a modification of the Herfindahl index to reflect relatedness in a more sensitive way by directly incorporating a component measuring heterogeneity in product input shares.[4] A similar development occurred with another measure familiar in the industrial concentration context, the entropy measure.[5] As used in the diversification literature, the entropy measure (Jacquemin and Berry, 1979; Palepu, 1985; Hoskisson, et. al, 1993) assesses aggregate relatedness by computing total diversification at the four-digit SIC level and then subtracting diversification computed at the two-digit SIC level, resulting in a related component or 'related entropy' that is based on the proportion of businesses that share the same two-digit class (Jacquemin and Berry, 1979). Like other methods, the entropy measure contains an explicit mechanism that gives greater weight to more significant activities.

measures considered here. Nevertheless, most measures implicitly include product counts but make the weighting issue the predominant consideration.

[4] Gollop and Monihan's (1991) heterogeneity component is derived by comparing vectors of 10 input shares to assess five-digit product diversification at the plant level. Inputs include production workers, other labor, fuel, electricity, purchased services, agricultural materials, mineral inputs, nondurable materials, durable materials, and capital.

[5] When individual shares are s_i, the standard calculation for entropy is $-\Sigma\, s_i \ln(s_i)$. This value is subtracted from one to give a concentration or diversification index qualitatively similar to the Herfindahl.

The concentric index (e.g., Caves, Porter, and Spence, 1980; Montgomery and Hariharan, 1991; Montgomery and Wernerfelt, 1988) contains a relatedness component based on the SIC system hierarchy. It is computed by first taking the product of shares of sales for each pair of businesses at the bottom level of the hierarchy and then multiplying that result by a digit representing the relationship between the two businesses in the SIC system. This digit takes on a value of 0 when the businesses are in the same three-digit category, a value of 1 when they belong to the same two-digit group but different three-digit groups, and a value of 2 when they are in different two-digit categories. Thus, like the Jacquemin and Berry (1979) measure based on entropy, concentric index relatedness is inferred directly from the hierarchical structure of the SIC system (See Table 1 for a summary of common diversification measures).

---Insert Table 1 about here -------------------------------------

In contrast to diversification measures which operate at the portfolio level, relatedness measures are designed to assess the relationship between two activity classifications and are therefore directly useful at the activity level. Relatedness measures are thus typically used *as a component* in a diversification construct in order to assess a portfolio-level strategy. In the case of the concentric index, for example, the measure is simply a weighting on intra-portfolio relatedness distances.[6]

Established diversification constructs typically depend on relatively crude measures of relatedness. This may account, in part, for challenges that they've received to construct validity, where it has been shown that relatedness effects are confounded with other features of the portfolio such as the number of businesses or size of the dominant business (Robins and Wiersema, 2003; Sambharya, 2000). For instance, the concentric index relatedness component offers only three possible values—0, 1, or 2 based on SIC hierarchy—certainly not a fine-grained assessment. Entropy relatedness is similarly limited. To compare the relatedness between two activities using the implicit two-digit versus four-digit

[6] As will be seen below, our approach uses weighting in development of the relatedness measure itself by assigning greater weights to activity pairs that constitute a greater portion of a firm's output.

diversification in the entropy measure, one would simply ask whether the activities share the same two-digit class. Answers are either yes (1) or no (0).

Importantly, relatedness components in standard diversification measures cannot effectively serve as stand-alone, general relatedness indices because the hierarchical structure of the SIC system does not represent an underlying relatedness scale.[7] Much of the SIC system reflects, for historical reasons, a broad logic of vertical structure and primary raw material. Thus, for example, functionally substitutable products made of steel, aluminum and plastic appear in different two-digit industries because of the underlying difference in primary feed stock. For other two-digit categories, and at finer classification levels, end-use plays a more significant conceptual role (electrical equipment, or apparel, for example). Ultimately, the fact that two four-digit industries share the same three-digit code and on up the line gives us no clear message about strategically significant relationships among activities. Relatedness simply cannot be reliably or directly inferred from the hierarchical structure of the SIC system (*cf.* Davis and Duhaime, 1992; Robins and Wiersema, 1995). A possible alternative to SIC hierarchy-based relatedness is the categorical method of relatedness identification based on researcher judgment (e.g. Rumelt, 1974). However, these methods apply a portfolio-level, not activity-level category designation. It is also open to possible bias due to the subjective nature of the relatedness judgments, which may lead different researchers to place the same firms in different diversification categories (Chatterjee and Blocher, 1992).

At the same time, fine-grained assessments of relatedness are crucial to empirically examining emerging strategy theory in areas such as the resource based view (Peteraf, 1993; Barney, 1991; Wernerfelt, 1984), organizational economics (e.g., Teece, 1980, 1982), and knowledge and capabilities (e.g., Winter, 2003, 1987; Helfat, 2000; Dosi, Nelson, and Winter, 2000; Teece, Pisano, and Shuen, 1997; Grant, 1996; Kogut and Zander, 1992). Empirical examination of these theoretical views requires the researcher to assess the degree of overlap, knowledge, or relatedness between one firm activity and another. For example, knowledge- and relatedness-based theorizing is used in discussions of how firms

[7] Gollop and Monihan's (1991) approach is more general and fine-grained than other approaches in that it uses the Euclidean distance between input shares to compute its relatedness component. However, this approach is highly production-centric and may therefore fail to capture broader notions of strategic relatedness or managerial logic.

search for new market-entry opportunities that allow the firms to economize on existing resources and knowledge as they build new capabilities (Bryce, 2003; Coff, 1999; Silverman, 1999; Teece, 1980, 1982); how capability evolution is built on sequences of decisions that are made in the context of resources already in hand (Helfat and Raubitschek, 2000; Helfat and Lieberman, 2002); or how the ability (owing to relatedness) to share firm-specific resources results in higher levels of firm performance (Teece, 1982; Petaraf, 1993; Mahoney and Pandian, 1992; Teece, Pisano, and Shuen, 1997).

Recent research requiring relatedness measures typically identifies a class of resources to be leveraged, develops a specific relatedness measure for the purpose, and then shows how firm expansion choices are related to that class, whether patents (Silverman, 1999), technology flows (Robins and Wiersema, 1995), human resource categories (Chang, 1992; Farjoun, 1990, 1994; Coff, 1999), or other areas. However, the identification of particular resource classes as the source of expansion may gain the power of specificity at the cost of generality, since the resources on which economies of scope are based can change with the firm or industry context in question. A summary of recent developments in relatedness measures is included in Table 2.

---Insert Table 2 about here --------------------------------------

3. Issues in developing a general relatedness index

The foregoing discussion suggests that a general measure of relatedness should meet at least the following four criteria.[8] The index should

(1) capture economies of scope in the large sense, without bias toward a particular economizing dimension;
(2) allow the underlying sources of such economies to vary among industries;
(3) forego reliance on the hierarchy of the SIC system for determining relatedness; and

[8] These criteria form an important subclass to the criteria for a diversification measure proposed by Gollop and Monihan (1991). Whereas Gollop and Monihan's criteria address diversification in general and include a criterion for relatedness, the criteria suggested here are for the relatedness measure itself.

(4) accommodate managerial logics for relatedness that may exist outside of obvious economizing motivations.

The measure should also preferably be continuous, rather than categorical.

As discussed previously, our approach relies on the survivor principle, and thereby responds appropriately to criteria (1) and (2). This eliminates the need to specify the precise basis of relatedness between any two industries since relatedness is inferred from the extent to which companies choose to combine the industries in their portfolios. Criterion (3) is satisfied because the method assesses relatedness without in any way invoking the SIC's hierarchical structure. The SIC system is used here only to categorize activity at the most micro level. Criterion (4) is addressed implicitly by building the index on a census of all firms in the US manufacturing economy down to firms with three employees. The importance of criterion (4) is suggested by the work of Stimpert and Duhaime (1997) and Pehrsson (2006), who have shown that conceptualizations of relatedness are multidimensional by analyzing managers' answers to a series of questions about the relationships between their businesses. The Herfindahl and entropy measures are strongly correlated to product market conceptualizations of relatedness, but other conceptualizations—such as financial, or commodity relatedness—are inadequately measured by standard indices (Stimpert and Duhaime, 1997). If these other conceptualizations are systematically employed by managers in actual diversification moves (e.g. Prahalad and Bettis, 1986; Grant, 1988), our measure will reflect them.

Basing a measure of relatedness on actual diversification patterns raises several important issues, however, which must be resolved. First, just because two industries have been combined in a portfolio by some firm does not mean it is a useful combination or that it should significantly influence the relatedness measure. Some combinations result from managerial experimentation or they arise for other unexplainable and unsystematic reasons (*cf.* Richardson, 1972). Furthermore, such "accidents" are more likely to occur when there are more trials. An industry in which many firms are active is more likely to be the site of such an accidental juxtaposition with a second industry than one that is sparsely populated. This issue is addressed by noting that the key to harnessing the information content in diversification is to reliably detect when combinations of industries are occurring inside portfolios at rates greater than one would expect if diversification moves were made at random. Teece et al. (1994) use such an approach to develop a normalized measure of the frequency by which industrial activities are combined within diversified firms.

Second, just because two activities appear together in some firm does not mean they are significantly related. Some very large portfolios contain relatively insignificant operations that may relate only to other minor activities in the portfolio. This second issue is addressed by weighting the normalized dyadic frequencies by the extent to which the two activities are both important in the overall economic picture of the firm. If an activity is insignificant whenever it is combined with a particular other activity in a portfolio, the dyad representing the combination should receive relatively less weight.

Third, the fact that two activities are not found combined in a single firm at a particular time does not necessarily mean that scope economies are entirely absent or, certainly, that the particular combination should be left without a valuation in the relatedness measure. As suggested above there can be costs as well as benefits from combining two activities within the same firm. In some cases, the fact that two activities don't appear in a firm may not indicate that there is nothing to be achieved by combining, but rather that the market provides a relatively effective means of combining relative to doing so within the firm (Teece 1980). The balance of costs and benefits may change over time, especially because some firms are gradually extending the scope of their capabilities. Our measure includes a provision that fills in the relatedness picture in cases where the direct evidence of actual joint participation is entirely absent. This involves creating a network representation of the weighted relatedness distances between industry nodes and computing the shortest path score between nodes. This procedure yields relatedness scores based on proximity in the network for activities that are not combined in any firm, and may imply increased estimates of the relatedness in cases where there is such joint participation, but the number of firms displaying it is relatively small.

Our data is drawn from the Longitudinal Research Database (LRD) at the Center for Economic Studies (CES) at the U.S. Census Bureau. The LRD represents the most detailed and extensive body of data on the productive inputs and outputs of U.S. manufacturing establishments (plants). The LRD is utilized instead of other possible alternatives for two basic reasons: (1) The LRD contains reliable information at the four-digit SIC level for all the activities in which firms actually engage, and (2) it provides a measure of the share of value-added produced by each firm in each four-digit product category, which supplies a measure of economic value that can be used to weight dyad counts for their importance to the firm. Of course, finer levels of classification exist in the SIC system, such as five-digit and even seven-digit codes. These codes are less commonly known to non-CES users, however, and computational complexity makes their use for the index difficult. The index relies on four-digit SIC codes insofar as it treats them as meaningful categories of activity, but it does not rely on the hierarchical structure or other relatedness approaches that could be extracted from the codes themselves. The data also has the distinct advantage of supplying a census rather than a sample of firms; operating data on all multi-unit firms that appeared in the 1987 Census of Manufactures (SIC 2000-3999) is included.[9]

3. Construction of the Index

Step 1. Following Teece et al. (1994), take industries two at a time and count the number of multi-industry firms operating in both industries. To be explicit, let $C_{ik} = 1$ if corporation *k* is active in industry *i,* and 0 otherwise. The number of corporations active in industry *i* is $n_i = \sum_k C_{ik}$, and the number of corporations active in industry *i* and *j* is $J_{ij} = \sum_k C_{ik} C_{jk}$. Raw counts of the number of firms

[9] A firm is defined as multi-unit when it operates two or more establishments with different primary four-digit SIC classifications. Excluded from the analysis are industries classified as "not elsewhere classified (n.e.c.)"—typically industry codes ending with a "9." These industries are "catch-all" categories containing a menagerie of products. In some cases, products are difficult to classify within alternative categories; in other cases, they are misclassified. Including n.e.c. industries in the analysis could bias the index because the network optimization process would likely produce pathways through at least some of these industries, creating relatedness scores that are potentially spurious.

operating in each industry dyad, however, cannot be taken directly as a measure of relatedness. Activities must be present at a rate *greater* than what one would expect if corporate diversification decisions were made at random. Although J_{ij} increases with the relatedness of i and j, it also increases with n_i and n_j, the number of firms operating in each industry of the dyad. Therefore, J_{ij} must be adjusted for the number of firms that would appear in the dyad under the null hypothesis that industries are assigned to firms at random (*cf.* Teece, et al. 1994).

To operationalize the null hypothesis, the distribution of J_{ij} must be derived. For now, call this random variable x_{ij}.[10] Under the null, the n_i firms operating in industry i are simply one random sample from the population of K multi-industry firms. Now draw another sample of size n_j and observe x, or the number of industries that were also in the n_i sample. The number of ways of selecting x firms to fill x positions in sample n_j is equivalent to the number of ways of selecting x from a total of n_i firms, or $\binom{n_i}{x}$.[11] The number of ways of selecting firms not receiving assignment to industry i for the remaining $(n_j - x)$ positions in the n_j sample is equivalent to the number of ways of selecting $(n_j - x)$ from a possible $(K - n_i)$ firms, or $\binom{K-n_i}{n_j-x}$. Then the number of possible permutations of the n_j sample is the number of ways of combining a set of x firms assigned to industry i (n_i) multiplied by $(n_j - x)$ firms not assigned to industry i, or $\binom{n_i}{x}\binom{K-n_i}{n_j-x}$.[12] The number of different samples of size n_j that can be drawn from K is

[10] Teece et al. (1994) identify the distribution, but they do not derive it in their paper. We found it necessary to derive the distribution in order to check what turned out to be minor typos in the original publication. Because doing so clarifies the set-up of the problem, we include the brief exposition here. The article is reprinted, with most if not all of the errors corrected, in Langlois, et. al., eds. (2003).

[11] $\binom{n_i}{x}$, or $C_x^{n_i}$ is the number of combinations*,* or subsets, of size x that can be formed from n_i objects and is computed as $\frac{n_i!}{x!(n_i - x)!}$.

[12] Since sample n_j was fixed as the number of firms operating in industry j, firms assigned to industry i in this quantity are *de facto* also assigned to industry j.

$\begin{pmatrix} K \\ \overline{n_j} \end{pmatrix}$. The number of possible permutations of the n_j sample divided by the number of ways of choosing a sample of size n_j is the probability that x firms from population K are assigned to both industry i and industry j. Thus, given joint participation of size x in two industries of size n_i and n_j , the number X_{ij} of corporations active in both industry i and industry j is a hypergeometric random variable

$$P\left[X_{ij} = x\right] = \frac{\begin{pmatrix} n_i \\ x \end{pmatrix}\begin{pmatrix} K - n_i \\ n_j - x \end{pmatrix}}{\begin{pmatrix} K \\ n_j \end{pmatrix}}. \tag{1}$$

Calculation in terms of factorials will serve to verify that the reversal of indices i and j has no effect on the value, so the apparent asymmetry in (1) is superficial. The mean of X_{ij} is

$$\mu_{ij} = E(X_{ij}) = \frac{n_i n_j}{K} \tag{2}$$

The variance of X_{ij} is

$$\sigma_{ij}^2 = \mu_{ij}\left(1 - \frac{n_i}{K}\right)\left(\frac{K - n_j}{K - 1}\right).^{13} \tag{3}$$

When the difference between J_{ij} and the expected value of the random variable x_{ij} is positive and large, it indicates systematic diversification by multi-industry firms into pairs of industries. The existence of pairs that are represented more frequently than suggested by the null necessarily implies a complementary set of relatively under-represented pairs. This does not imply some sort of negative relatedness, but only that

[13] Intuition for the mean of (1) is as follows. Assume that n_j firms in K have been assigned to industry j. Now randomly assign firms in K to industry i. The probability that any one firm receives an industry i assignment is $\frac{n_i}{K}$. Since there are n_j firms in K, each with probability $\frac{n_i}{K}$ of being assigned to industry i, the expected number of firms assigned to both industry i and industry j is $n_j\left(\frac{n_i}{K}\right)$. For further information on the hypergeometric distribution, see Feller (1957).

the incentives to participate in such pairs are weak relative to the stronger forces affecting the over-represented pairs. The difference between J_{ij} and the expected value of x_{ij} is standardized as

$$\tau_{ij} = \frac{J_{ij} - \mu_{ij}}{\sigma_{ij}} . \tag{4}$$

Step 2. Since Equation (4) is based on raw industry participation counts, it is a coarse measure of the extent to which activity combinations are economically important to the firm. The normalization process corrects for the frequency with which industry dyads occur across firms, but it does not reflect the economic importance of the dyad to the average firm operating in the dyad. If an industry dyad is responsible for only a small fraction of the economic value produced by each firm that participates in it, it hardly seems reasonable to accord this joint participation the same weight as other combinations that are more important to the firms involved. In a broadly diversified firm, two activities each delivering only 1 - 2 percent of the firm's value-added may be only weakly related, whereas two activities in a smaller firm that each deliver close to half of the value-added are likely related more strongly. If the pattern is consistent across all firms operating in two focal industries, then relatively lower or higher weights, as appropriate, should be assigned to the relatedness score of the dyad. This is what our index does.

The weight is determined by comparing for each dyad the relative proportions of total firm value-added that are attributable to each activity of the dyad. The minimum of these two value-added proportions is then selected for each firm and averaged across all firms operating in the dyad. The minimum proportion is selected because it represents an "upper bound" measure of how closely related the two industries could be when they appear together. If industry A, having a value-added proportion of 0.01, is combined with industry B, having a value-added proportion of 0.7, the 0.01 is selected to provide information on the importance of the dyad to that firm. In another firm with the same dyad, industry B could have the smaller proportion, in which case industry B's proportion would be selected to provide the information. These minimum proportions are then averaged across all firms operating in the dyad to create the dyad weight. The average S_{ij} produced by all firms operating in the dyad is

$$S_{ij}^{\min} = \frac{\sum_k \min_k [s_i, s_j] C_{ik} C_{jk}}{\sum_k C_{ik} C_{jk}}. \tag{5}$$

Scores in Equation (4) are then adjusted by the weights in Equation (5). Before weighting, the scores in (4) are converted to a distance matrix, a necessary setup for computing shortest path distances in Step 3. The distance matrix is computed by identifying the maximum τ_{ij} among the set of normalized scores, and subtracting all scores from this value. In the distance matrix, low cell values mean high relative relatedness and zero represents the most related dyad. All other values are positive. Following this transformation, cell values in the distance matrix are divided, not multiplied, by (5). After weighting by (5), the resulting matrix can be evaluated as a network in which the values in matrix cells are the distances between nodes *i* and *j*. The network is comprised of industry vertices connected by arcs having weight (length) inversely proportional to relatedness. Every pair of industries found together in a diversified firm has a corresponding arc-length in the network. Note however that, at this stage, only the *ij* pairs combined empirically are directly connected, all others remain unconnected. If indirect connections are considered – such as *i* to *k* and *k* to *j*, or longer chains – then we find that the network as a whole is connected with the exception of three minor cases that are strict isolates, SIC 2386, 2371, 3263. These three industries are dropped from further consideration.

Step 3. To be useful as a tool for determining relatedness for any expansion option facing the firm, the measure should supply scores for all possible industry combinations, including those that are not observed in the timeframe for which the measure is constructed. This issue is addressed by solving for the shortest path distance between every pair of nodes in the weighted distance matrix.[14] The method

[14] Computation of the shortest path through a network is a well-known problem and has a straightforward formal representation. Consider a network consisting of industry node (vertices) set *V* and arc (edge) set *E*. Each edge $e \in E$ has cost $c_e = \delta_{ij}^{w}$, which is the weighted distance between industry nodes $v_i, v_j \in V$. Consider one pair of nodes v_1 and v_k. The total cost of a path $p \in P = v_1 e_1 v_2 e_2 \ldots v_{k-1} e_{k-1} v_k,\ v_i \in V,\ e_i \in E$ is the sum of the costs of the

produces a distance measure for dyads that are not directly connected in the network, and it substitutes a shortest path distance for a direct link between two industries when the path distance is shorter than the direct distance. The substitution also produces a measure that is, by construction, a legitimate “distance” in the mathematical sense underlying the concept of a metric space, namely, that the resulting relatedness scores satisfy the triangle inequality: $d(x,y)+d(y,z)\geq d(z,x)$, where $d(x,y)$ is the distance between x and y (Takayama, 1985). To illustrate, consider Figure 1, a representative network, where letters represent industry nodes and lines represent the arc-length distance between industries (the shorter the arc, the more related the industry). In the simple network of Figure 1, industry node *A* and node *E* are not connected directly, but node *E* can be reached along the shortest path *ABE.* The distance represented by *ABE* becomes the computed relatedness distance for *AE.* The shortest path calculation could also lead to the replacement of existing distances based on actual joint participation with shorter ones based on stronger indirect connections.

--Insert Figure 1 about here --------------------------------------

To complete construction of the index, the weighted distance matrix, which is now filled with shortest path scores, is converted to a similarities matrix, where the greatest values rather than the lowest values represent the highest relatedness. This is done simply by subtracting each computed path length score from the maximum computed path length, which implicitly sets the least related dyad to a value of zero and the most related dyad to some positive value. Following the similarities transformation, index scores are further transformed in two ways. In the first, the similarities score is standardized by subtracting the mean of the distribution from each value and dividing by the standard deviation. These

edges on this path $c=\sum_{i=1}^{k-1}c_{e_i}$. The problem is to find the path *P* that begins at v_1 and ends at v_k such that *c* is a minimum.

scores are distributed approximately normally but the distribution has a long, left tail, implying that there are a number of dyads with very low relatedness. Normalized values, or z-scores, range from a low of -7.00 to a high of 3.51 standard deviations from the mean. In the interest of interpretability, the relatedness scores are also transformed into a value that represents the cumulative area under the distribution and ranges between 0 and 100. Here the scores may be interpreted as a percentile. An index score of 70 implies that 70 percent of industry dyads are less related than the focal score, while 30 percent are more related. Plots of the distribution of all normalized (not percentile) dyad relatedness index scores are shown in Figure 2. Note that Figure 2 represents only the distribution of raw scores between every industry pair. It does not represent the relatedness scores of industries inside firm portfolios.

--Insert Figure 2 about here --------------------------------------

A few examples of scores illustrate the ability of the index to capture relatedness relationships among industries; the examples also supply face validity. First, illustrating relatively low relatedness, SIC 3264, “Porcelain Electrical Supplies,” and SIC 2421, “Sawmills and Planing Mills,” score near the zero percentile of relatedness (0.25 percentile) with a z-score of -4.69, suggesting that these activities share little in common.[15] The relatedness here squares with what intuition might suggest; the advantage of the index is that it provides a precise relative measure in comparison to other dyads. The two most unrelated industries are SIC 2097, “Ice,” and SIC 2397, “Schiffli Machine Embroidery,” with a z-score of -7.0. In contrast, the two most related industries, receiving a z-score of 3.51 and a percentile rank of 100, are SIC 2131, “Tobacco, Chewing and Smoking and SIC 2141, “Tobacco Stemming and Redrying.” The index seems to confirm intuition for these pairs of industries.

[15] These two industries indicate the lowest relatedness outside of dyads that include SIC 2397, “Schiffli Machine Embroideries.” The latter SIC code accounts for all z-scores in a range lower than -4.69, down to -7.0. Apparently, this industry is less related to a higher number of dyads than all other industries. Industry 2397 produces embroidered textile products using a Schiffli embroidery machine which was invented by Isaac Groebli of Switzerland in the late 1800s. The machine utilizes a continuously threaded needle and a shuttle containing thread. The shuttle looks similar to the hull of a sailboat. Thus, the machine garnered the name "Schiffli," which means "little boat" in the Swiss German language.

The index identifies numerous examples of very high levels of relatedness between pairs of industries that are different at the two-digit level within manufacturing. Hierarchy-based relatedness methods typically consider two-digit differences to be unrelated. As just one example, consider SIC 2951, "Paving Mixtures and Blocks," and SIC 3273, "Concrete, Ready-Mixed." The percentile rank here near 100 (z-score 3.07) is not surprising given the category descriptions, yet none of the typical approaches to SIC hierarchy-based relatedness would have detected this relationship. A more interesting example is the percentile relatedness near 100 (z-score 3.04) between SIC 2542, "Metal Partitions and Fixtures" and SIC 3581, "Automatic Vending Machines." This high index score suggests that complementarities may exist in combining what appear on the surface to be disparate activities. Digging a bit deeper, it seems clear that knowledge about how to manufacture or distribute metal frames could be made applicable to manufacturing or distributing the frames on vending machines.

Consider an example of using the index to predict an expansion move. In 2003, Energizer Holdings, Inc., a battery manufacturer, acquired Schick-Wilkinson Sword, a safety razor manufacturer, to diversify its product line ("Energizer acquires Schick," 2003). While the logic for this move is not immediately evident, Pat Mulcahy, chief executive officer of Energizer, supplies the following rationale:

> Schick-Wilkinson Sword is an attractive business in a category with dollar sales growth and stable margins that leverages our core competencies. . . . Energizer and Schick are very compatible, with many common customers, and similar distribution channels, high speed manufacturing and product innovation capabilities, and corporate cultures ("Energizer acquires Schick," 2003).

The CEO apparently used several resource categories and a complex logic in evaluating the relatedness between these two opportunities. If the CEO's assessment is accurate, knowledge overlap exists between razors and batteries because they serve common customers, have similar distribution channels, use manufacturing technology with significant similarity, and share similar product innovation and corporate cultures. Use of any one of these resource categories to identify this opportunity may or may not have been successful. Thus, an important question is whether the general index developed here could have detected *a priori* this sort of non-obvious opportunity. The most likely classification for the batteries manufactured by Energizer Holdings, Inc. and the safety razors manufactured by Schick-Wilkinson Sword are SIC 3691, "Storage Batteries," and SIC 3421, "Cutlery," which includes safety razors, respectively. Although Census lumps alkaline cell batteries of the type manufactured by Energizer together with automobile lead acid storage batteries and also other types (which dilutes the focus of the category), and also lumps razor blades, scissors, and shears together with safety razors, the relatedness percentile between these industries is still 62 (z-score 0.31), a stronger relatedness than average, and stronger than one might expect *a priori*. The index uncovers relatedness between what appear to be unrelated industries, and yet the findings are consistent with a managerial logic that suggests the presence of complementarities in razors and batteries.

These examples indicate that the index is doing at least part of what it was intended to do: Uncovering relatedness between pairs of industries, independent of the specific source of economies of scope.

4. Test of Predictive Validity

The general relatedness index developed here is intended to be precisely that—general. The number of its potential applications is very broad. We select here just one specific and conservative application—an empirical context for which theory suggests that relatedness effects are likely to be particularly strong: the entry mode choice. This is a particular use of the general index which, as we argued above, captures multiple underlying bases of relatedness. Here we argue that an influential factor in the decision about whether to build or acquire as a mode of entry is the extent to which the firm holds knowledge that is specific enough to qualify it as the creator of a production function in a target market. This type of knowledge specificity in production is among the sets of possibilities for what may be causing firms to jointly participate in industries, and the index should therefore reflect it, even if only weakly among the alternative possibilities that motivate activity combinations. As applied here, the index is a proxy for shared, specific knowledge, where higher index scores indicate that more specific knowledge is common to two activities in view.

We establish in this brief exercise that the index has substantial predictive validity (understood as the degree to which a measure of a concept shows the expected statistical relationship with some recognized outcome (Lubatkin et al. 1993: 436)). This cross-industry exercise also illustrates the common situation in which the resources underlying "relatedness" cannot be consistently classified for all industries—requiring the kind of general index developed here. The results demonstrate the index's usefulness as a general empirical tool, its predictive validity, and its advantages over alternative relatedness constructs based on SIC hierarchy. The results also validate the conceptual adjustments made to the Teece et al (1994) measure, which in its original form does not turn up as significant in our tests.

Theory

A firm's choice about mode of entry cannot be made independent of the characteristics of knowledge in hand. Since a new establishment is presumptively an establishment without a production function, it is an asset that is likely to be attractive only to an investor capable of supervising the creation of the appropriate production function. The obvious candidate for the role is a firm that already possesses a similar production function in a similar establishment. The requisite coordinating information for productive activity is partly imported into the establishment in the skill sets and mental models of personnel, partly accumulated locally through learning-by-doing (with early productive efforts likely to yield more learning than product), and partly embodied in fragmentary form in the establishment itself. By contrast, a functioning establishment that has been "previously owned" when acquired is a real asset generating cash flows that can be reasonably estimated on the basis of past experience. Likewise, the firm may be the only entity qualified to build its new plant if this requires careful replication of highly technical knowledge and routines (Winter and Szulanski, 2001). When Intel Corporation must build a new fabrication facility, for example, the company does not go shopping for the facility on the open market. Instead it builds the facility using its own specific, highly technical knowledge. Similarly, Helfat and Lieberman (2002) argue that the greater the required resources and capabilities that firms possess prior to entry, the more likely they are to use internal growth, or build modes. Early work examining the choice of entry mode also suggested a positive correlation between the relatedness of existing activities and the target industry (Yip, 1982). In contrast, when firms seek to leverage some pre-entry resources and capabilities but lack other critical, especially specific resources, they are more likely to choose entry by acquisition. Of course, if the acquiree is sufficiently distant from the acquiror's knowledge base, or if the knowledge required to run the target is specialized, successful entry through acquisition may be difficult. Nevertheless, acquisition may be the only option when the firm lacks the specific knowledge that would make it an effective builder.[16]

The maintained hypothesis underlying our test of predictive validity may be summarized as follows: *Expanding firms that possess specific knowledge related to a focal market will typically choose to enter by building, rather than acquiring, a new establishment*.

Data and Methods

The sample for the analysis includes all establishments from the LRD that were built or acquired by a continuing firm between the 1987 and the 1992 economic censuses. The plant must have been in a four-digit industry in 1992 in which the owning firm did not participate in 1987. The number of such establishments is 4,721. However, due to missing values for select covariates (e.g., industry R&D expenditures), the number of establishments included in the regression analysis is reduced to 1,706.[17] The choice of entry mode is modeled as a dichotomous variable where 1 is entry by build, and 0 is entry by acquisition. A probit specification is utilized for the two-period panel. All manufacturing firms operating in 1987 that by 1992 had entered a new (four-digit) industry are considered. Theoretical and control variables are listed below.

[16] Some plants may be built on behalf of the focal firm by specialist engineering firms who bring technical knowledge to get a "turnkey" plant up and running (e.g. Arora, Fosfuri and Gambardella, 2001). This phenomenon represents a kind of intermediate category between build and acquisition. To the extent that such instances exist in our data, they are coded as build. However, since specialist firms allow focal firms to build plants in industries that are actually further from their domain of expertise, the presence of these instances in our data will work against our results and thus makes our test more conservative.

[17] R&D intensity is calculated at the industry level based on COMPUSTAT (See Appendix). R&D-intensive industries are likely to require the development of specialized resources for effective competition. Holders of specialized resources are more likely to enter by build. We thus view this variable as an important control on the findings. Running the analysis without the R&D intensity variable does not qualitatively change the results but clearly increases the number of observations in the regressions. Coefficients on relatedness and other theoretical variables were, as a result, more significant in these runs. We do not include those results here.

Relatedness. Relatedness is measured in three different ways for comparison. The first measure is a naïve, two-digit measure, which is coded 1 if in 1987 the entering firm owned establishments operating in the same two-digit industry as the 1992 entered industry, and coded 0 otherwise. Inclusion of this variable supplies a basic test of whether the relatedness component of entropy or the concentric measure is able to distinguish entry mode, since the relatedness in these measures is based on shared hierarchy within the SIC system. The second approach is the Teece et al. (1994) measure identified by Equation (4) above, which provides a basic test of whether the adjustments made to convert the measure into a general relatedness index are effective. The third measure is the general relatedness index. Each of these measures approximates the relatedness to the target industry of the most related other industry in the portfolio. A positive sign is expected on relatedness coefficients, indicating that relatedness increases the probability of a build choice.

Coherence. Coherence (Teece et al., 1994) is defined as the employee-weighted average value of the relatedness of activity dyads on the maximum spanning tree of a firm's activity portfolio. In essence, it is the average relatedness of each industry linked to its closest other industry in the portfolio. In that regard it is in one sense a portfolio-level, related diversification measure. Knowledge-based theorizing suggests that firms enjoying very tight coherence in their activity set would be more likely to possess and deploy specific knowledge in entry decisions. The converse is also true. Less coherent firms are more likely to deploy general knowledge, such as in acquisition (Montgomery and Wernerfelt, 1988). Thus, exclusion of this variable could confound the independent effects of knowledge specificity to the target industry.

Experience. The length of experience in a general area is coarsely defined as the number of years of operating experience in the two-digit industry in which the target four-digit industry is found. Although we limit the sample to four-digit industries in which the firm has never operated, the firm may have operated in the two-digit class of that industry. We sum years of experience in the two-digit industry since the 1963 Census. This provides a further control on the relatedness variable since it proxies the knowledge the firm may have already acquired through accumulated experience in activities close to the target.

Following standard approaches to modeling entry (Geroski, 1991), controls for firm size and industry structure (industry growth, concentration, asset intensity, profitability, build propensity, and R&D intensity) are also included (See Appendix for a detailed description). Pearson correlation coefficients for all variables are shown in Table 3.

--Insert Table 3 about here -------------------------------------

Results

Results of the probit regression analyses are shown in Table 4. Strong support is found for effectiveness of the general index as a predictor of choice of mode of entry—it is highly significant at $p<0.001$. Model 4, which incorporates the general relatedness index, performed the best overall.

--Insert Table 4 about here -------------------------------------

Only the general relatedness index shows statistical significance in this analysis. Other indicators, including the two-digit measure and the Teece et al (1994) measure of Equation (4), were not significant. As it pertains to the index, this result clearly indicates that the method of value-added conditioning and shortest path search contributes important information to the task of assessing relatedness. The general index differentiates between situations favoring build and acquire entry modes whereas the other measures do not – at least, not in the presence of the control variable we employ.

Control variables coherence, firm size, and the proportion of new (startup) firms that build versus acquire in the target industry (a measure of propensity to build which may reflect complexity of knowledge in the particular industry) were highly significant in the expected direction. Experience, pre-entry industry growth rate, asset intensity, average plant profitability, and R&D intensity were significant but less so, and experience and concentration were marginally or not significant, depending on the model. Findings on control variables suggest the models effectively control for the influence of relatedness on mode of entry.

Predictive validity summary

The net result of the predictive validity assessment is that the proposed relatedness measure shows the expected statistical relationship with the recognized outcome of building rather than acquiring. This result is obtained with the general index and is not based on whether a firm is entering from an industry that shares the same two-digit class, nor is it obtained with the construct of Equation (4) alone.

5. Discussion

This section briefly discusses some limitations of the general index and also some possible applications of the index in corporate strategy, and longitudinal research.

Limitations

The general relatedness index should be a useful tool for assessing inter-industry relatedness in virtually any context requiring such a measure. One limitation, however, is that the version of the index developed here only provides relatedness scores for industries in the Manufacturing sector. Additionally, the index is based on multi-product organization and diversification as it existed in 1987. As industries change and technology develops, the relatedness relationships between industries may also change. Nevertheless, given the methodology of statistical normalization, value-added weighting and averaging, and shortest path search, the relationships developed here are expected to be stable and durable, making the index useful for general questions performed on data existing before or after the 1987 construction year. An additional virtue of the index is that it need not be computed anew each time it is used. Its strong empirical base—all diversified firms in the US economy of any size—makes repeated construction costly and difficult.[18] Hence, the authors are making the index generally available to academic researchers. Needless to say, however, an effort to recalculate the index on the basis of more recent data would be welcome, and would afford some direct insight into the stability of the patterns captured by it.

Applications

The potential applications of the index are broad, but the index holds particular promise in studies of firm expansion and diversification, where it offers new empirical content for theoretical logic based on the resource-based view of the firm. Some specific areas of promise are identified below.

Related vs. Unrelated Diversification. Using the index, it is possible to construct detailed profiles of firm portfolios and fine-grained measures of relative relatedness among all industrial activities in each portfolio. Examination of intra-portfolio relationships at a micro level with a more fine-grained relatedness measure has the potential to provide additional insights into familiar questions about the links between diversification strategy and performance. For example, aside from the related-unrelated dichotomy, are some particular portfolio configurations more advantageous to performance than others? Similarly, how does the positioning of particular activities inside the portfolio impact the performance of those activities? Do activities that are more central in the intraportfolio relatedness "network" perform better (or persist longer) than those that lie on the periphery? Since certain activities come to share knowledge, capabilities, and resources by virtue of their similarities inside the firm, one would expect positive performance effects to exist between closely related industrial activities inside the portfolio. Activities that are deeply embedded in a relatedness sense within the portfolio of an evolving firm are more likely to experience spillovers in knowledge, resources, and capabilities from multiple sources (Tsai, 2000), potentially leading to positive performance effects.

Longitudinal Strategy Research. Emerging strategic theory draws heavily on Penrose's *Theory of the Growth of the Firm* (1959) to explain the direction of expansion, the development of capabilities, and

[18] We have constructed the index using only publicly available firms available in the Compustat Segment files but found generally weak correlation to the general index.

the role of knowledge in the growth of the firm. Fundamentally, such theories are about firm growth and therefore, in a diversified firm, require longitudinal assessments of market entry choices. Yet, perhaps surprisingly, there are a limited number of empirical studies in the literature that take this perspective. No doubt the lack of good tools for assessing patterns of longitudinal expansion choices has been a prime contributor to the deficit. A number of interesting questions remain to be explored. For example, do firms that make repeated series of short leaps into new markets outperform or underperform firms that make longer leaps? What determines which opportunities can profitably be seized by which firms? Do firms that appear to repeatedly leverage a core strength in some industry into other related industries perform better or worse than firms that operate from multiple capability platforms simultaneously? How do capabilities evolve in the multiunit firm as a function of market entry? Does the rate of development in these capabilities have anything to do with performance? We expect that, with the help of the general index introduced here, researchers will be able to tackle longitudinal questions such as these with renewed vigor.

REFERENCES

Arora, A., A. Fosfuri and A. Gambardella. 2001. Specialized technology suppliers, international spillovers and investment: evidence from the chemical industry. *Journal of Development Economics* 65(1):31-55.

Barney, J.B. 1991. Firm resources and sustained competitive advantage. *Journal of Management* 17(1): 99-120.

Berry, C.H. 1971. Corporate Growth and Diversification. *Journal of Law and Economics*, 14(2): 371-383.

Bryce, D.J. 2003. Firm knowledge, stepping stones, and the evolution of capabilities. Doctoral dissertation. University of Pennsylvania.

Caves, R.E., M. Porter and A.M. Spence. 1980. *Competition in the open economy: A model applied to Canada*. Cambridge, Mass.: Harvard University Press.

Chandler, A.D. 1962. *Strategy and Structure.* MIT Press, Cambridge, Mass.

Chang, S.J. 1992. A knowledge-based perspective on corporate growth: Entry, exit, and economic performance during 1981-1989. Ph.D. dissertation. University of Pennsylvania.

Chatterjee, S. and J.D. Blocher. 1992. Measurement of firm diversification: Is it robust? *Academy of Management Journal* 35(4): 874-888.

Coase, R. H. 1937. The Nature of the Firm. 4 *Economica* n.s. 386-405.

Coase, R. H. 1991. Nobel Lecture: The institutional structure of production. Reprinted in O. Williamson and S.G. Winter (eds.), *The Nature of the Firm*. New York: Oxford University Press.

Coff, R.W. 1999. How buyers cope with uncertainty when acquiring firms in knowledge-intensive industries: Caveat emptor. *Organization Science* 10(2), 144-161.

Davis, R. and I. Duhaime. 1992. Diversification, vertical integration, and industry analysis: New perspectives and measurement. *Strategic Management Journal*, 13 (7): 511-524.

Dosi, G., R.R. Nelson and S.G. Winter. 2000. *The nature and dynamics of organizational capabilities.* Oxford University Press: Oxford.

“Energizer Holdings, Inc. Acquires Schick-Wilkinson Sword.” 2003. January 21. *PRNewswire-FirstCall*.

Farjoun, M. 1990. Beyond industry boundaries: Human expertise, diversification, and related industry group. Ph.D. dissertation, Northwestern University.

Farjoun, M. 1994. Beyond industry boundaries: Human expertise, diversification and resource-related industry groups. *Organization Science,* 5(2): 185-199.

Feller, W. 1957. *An Introduction to Probability Theory and Its Applications, Vol. I* (2nd. ed.). Wiley: New York.

Geroski, P.A. 1991. *Market Dynamics and Entry*. Basil Blackwell: Oxford.

Gollop, F.M. and J.L. Monihan, 1991. A generalized index of diversification: Trends in U.S. manufacturing. *The Review of Economics and Statistics* 73(2): 318-330.

Gort, M., 1962. *Diversification and Integration in American Industry*. Princeton University Press, Princeton, NJ.

Grant, R.M. 1988. On 'dominant logic', relatedness and the link between diversity and performance. *Strategic Management Journal* 9: 639-642.

Grant, R.M. 1996. Toward a knowledge-based theory of the firm. *Strategic Management Journal* Winter: 109-122.

Helfat, C. E. 2000. Guest editor's introduction to the special issue: The evolution of firm capabilities. *Strategic Management Journal* 21(10-11): 955-959.

Helfat, C. E. and L. Raubitschek. 2000. Product sequencing: Co-evolution of knowledge, capabilities and products. *Strategic Management Journal* 21(10-11):961-979.

Helfat, C. E. and M. J. Lieberman, 2002. The birth of capabilities: Market entry and the importance of pre-history. *Industrial and Corporate Change* 11:725-760.

Hoskisson, R.E., M.A. Hitt, R.A. Johnson, and D. Moesel. 1993. Construct validity of an objective (entropy) categorical measure of diversification strategy. *Strategic Management Journal* 14(3): 215-234.

Jacquemin, A.P. and C.H. Berry. 1979. Entropy measure of diversification and corporate growth. *Journal of Industrial Economics*, 27(4): 359-369.

Kogut, B. and U. Zander. 1992. Knowledge of the firm, combinative capabilities, and the replication of technology. *Organization Science* 3:383-397.

Lubatkin, M., H. Merchant and N. Srinivasan. 1993. Construct validity of some unweighted product-count diversification measures. *Strategic Management Journal* 14(6): 433-449.

Mahoney, J.T. and J.R. Pandian. 1992. The resource-based view within the conversation of strategic management. *Strategic Management Journal* 14(6).

Montgomery, C.A. 1979. Diversification, market structure and firm performance: An extension of Rumelt's model. Ph.D. dissertation, Purdue University.

Montgomery, C.A. and S. Hariharan, 1991. Diversified expansion by large established firms. *Journal of Economic Behavior and Organization* 15:71-89.

Montgomery, C.A. and B. Wernerfelt. 1988. Diversification, Ricardian rents, and Tobin's *q*. *RAND Journal of Economics* 19(4):623-632.

Palepu, K. 1985. Diversification strategy, profit performance and the entropy measure. *Strategic Management Journal* 6:239-255.

Panzar, J. and Willig, R. 1981. Economies of scope. *American Economic Review* 71: 268-272.

Pehrsson, A. 2006. Business relatedness and performance: a study of managerial perceptions. *Strategic Management Journal* 27(3): 265-282.

Penrose, E. 1959. *The Theory of the Growth of the Firm.* New York: John Wiley.

Peteraf, M.A. 1993. The cornerstones of competitive advantage: A resource-based view. *Strategic Management Journal* 14, pp. 179-192.

Prahalad, C.K. and R.A. Bettis. 1986. The dominant logic: a new linkage between diversity and performance. *Strategic Management Journal* 7: 485-501.

Ramanujam, V. and P. Varadarajan. 1989. Research on corporate diversification: A synthesis. *Strategic Management Journal* 10(6): 523-551.

Richardson, G.B. 1972. The Organisation of Industry. *The Economic Journal* 82(327):883-896.

Robins, J. and M.F. Wiersema. 1995. A resource-based approach to the multi-business firm: Empirical analysis of portfolio interrelationships and corporate financial performance. *Strategic Management Journal* 16 (4): 277-299.

Robins, J.A. and M.F. Wiersema. 2003. The measurement of corporate portfolio strategy: Analysis of the content validity of related diversification indexes. *Strategic Management Journal* 24: 39-59.

Rumelt, R.P. 1974. *Strategy, structure and economic performance.* Harvard University Press: Boston, Mass.

Rumelt, R.P. 1982. Diversification strategy and profitability. *Strategic Management Journal* 3(4):359-369.

Sambharya, R.B. 2000. Assessing the construct validity of strategic and SIC-based measures of corporate diversification. *British Journal of Management* 11: 163-173.

Scherer, F.M. 1982. Intra-industry technology flows in the U.S. *Research Policy* 11:227-245.

Silverman, B.S. 1996. Technical assets and the logic of corporate diversification. Doctoral Thesis. University of California, Berkeley.

Silverman, B.S. 1999. Technological resources and the direction of corporate diversification: Toward an integration of the resource-based view and transaction cost economics. *Management Science* 45(8):1109-1124.

Stigler, G.J. 1968. *The organization of industry*. Irwin: Homewood, Ill.

Stimpert, J.L. and I.M. Duhaime. 1997. In the eyes of the beholder: Conceptualizations of relatedness held by the managers of large diversified firms. *Strategic Management Journal* 18(2): 111-125.

Takayama, A. 1985. *Mathematical economics*. Cambridge University Press: New York (2nd Edition).

Teece, D., R. Rumelt, G. Dosi, and S. Winter. 1994. Understanding corporate coherence: Theory and evidence. *Journal of Economic Behavior and Organization* 23:1-30. Reprinted (with corrections) in *Alternative Theories of the Firm, Vol. II*, R. N. Langlois, T. F-L. Yu and P. Robertson, eds. Cheltenham, UK: Elgar, 2002.

Teece, D.J., G. Pisano, and A. Shuen. 1997. Dynamic Capabilities and Strategic Management. *Strategic Management Journal* 18(7): 509-533.

Teece, David J. 1980. Economics of scope and the scope of the enterprise. *Journal of Economic Behavior and Organization* 1, 223-247.

Teece, David J. 1982. Toward an economic theory of the multiproduct firm. *Journal of Economic Behavior and Organization* 3, 39-63.

Tsai, W. 2000. Social capital, strategic relatedness and the formation of intraorganizational linkages. *Strategic Management Journal* 21:925-939.

Wernerfelt, B. 1984. A Resource-Based View of the Firm. *Strategic Management Journal* 5(2), pp. 171-180.

Williamson, O.E. 1985. *The Economic Institutions of Capitalism*. New York: The Free Press.

Winter, S.G. 1987. Knowledge and competence as strategic assets. In D.J. Teece (ed.), *The Competitive Challenge: Strategies for Industrial Innovation and Renewal.* Cambridge, Mass: Ballinger: 159-184.

Winter, S.G. 1988. On Coase, competence and the corporation. *Journal of Law, Economics and Organization* 4: 163-180.

Winter, S.G. 2003. Understanding dynamic capabilities. *Strategic Management Journal* 24(10): 991-995.

Winter, S.G. and G. Szulanski. 2001. Replication as strategy. *Organization Science* 12:730-743.

Wrigley, L. 1970. Divisional autonomy and diversification. Unpublished doctoral dissertation, Harvard Business School.

Yip, G. 1982. Diversification Entry: Internal Development versus Acquisition, *Strategic Management Journal,* 3:331-345.

APPENDIX

Size of the parent firm is computed as the natural log of the total value of shipments (TVS) for the firm across all its industry operations in 1987. We expect a negative sign on the coefficient of size since large firms are more likely to acquire given greater access to external financial resources (Chatterjee & Singh, 1991).

Pre-entry industry growth rate is a measure of industry attractiveness. It is measured in 1987 as the total industry growth in TVS since the 1982 economic census in order to capture the growth rate faced by firms at the beginning of the period under study (1987-1992). A rapidly growing industry is likely to attract firms from afar who are interested in investing even without the industry-specific resources. But because the industry is growing rapidly, there are likely to be few firms available for acquisition, especially at a price less than the future discounted rent stream. Thus, we expect that in rapidly growing industries, much of the growth is fueled by internal development by firms possessing the right resources and this provides a further control on relatedness.

Four-firm concentration ratio measures industry concentration for the four largest firms in the industry as the industry proportion of total value of shipments accounted for by these firms. We expect higher concentration ratios to be associated with oligopolistic rivalry conditions, larger average firm size and higher barriers to entry. If the largest firms control a significant portion of the capacity in the industry, then entering firms may need to acquire to gain a foothold—i.e. the sign on the coefficient is expected to be negative.

Industry asset intensity measures the capital requirements for entrants. It is calculated as the natural log of industry investments in plant & equipment in 1992. On the one hand, intensive capital requirements may suggest that large firms with deep pockets will tend to enter by acquisition. On the other hand, it may be the case that intensive capital requirements are the sorts that require specific knowledge—such as in highly technical industries requiring heavy expenditures in R&D. Thus, rationale for sign of either direction can be developed and we make no prediction about the sign of this variable.

Industry profitability is a measure of industry attractiveness determined as the average plant-level profitability in the industry, which is computed as value added (less labor) divided by total value of shipments (TVS) in 1992—conceptually the profit potential entrants can hope to earn per plant. We expect profitability to attract well-financed entrants who are conducting broad search for profitable opportunities. Thus, we expect that entrants will be induced to acquire in hopes of purchasing the cash flow stream as early as possible. This implies that the sign will be negative.

Industry build propensity is calculated as the ratio of new (startup) firms that build vs. acquire and is a relative measure of the extent to which entry by build is straightforward in the industry, perhaps owing to the particular technology required for success. For example, in some industries, knowledge required for successful production may be so general, and start-up costs so low, that firms nearly always build rather than acquire establishments, even when acquirable establishments are available. Using this measure as a control also serves as a proxy for other potentially unobserved determinants of the propensity to build in the industry and ensures that remaining effects of differences in build and acquire are owing to theoretical variables (and other controls). We expect a positive sign.

R&D intensity is the extent to which research and development (R&D) is a factor in a particular industry and is measured as average R&D expenditures over total revenues from COMPUSTAT for 1992 in each four-digit industry. Unfortunately, not all four-digit industries identified in the LRD are found in COMPUSTAT. When a four-digit value was not available, and where possible, we utilized the average R&D intensity at the three-digit level. Even after this adjustment, however, a number of establishments could not be matched on an R&D intensity score. This effectively reduced the set of industries analyzed to those in which R&D is a factor. R&D-intensive industries are likely to require the development of specialized resources for effective competition. Holders of specialized resources are more likely to enter by build. Therefore, we expect a positive sign on the coefficient.

Table 1: Select Diversification Measures

Measure	Mathematical Form	Empirical Base	Relatedness Component	Primary usage	Source
1. Herfindahl Index	$D = 1 - \sum_{i=1}^{n} s_i^{\,2}$ where n = number of activities in portfolio and *s* = each activities' share	Patterns of firm revenues within portfolio	None in standard measure. Gollop and Monihan (1991) insert Euclidean distances among product class input shares	Diversification research; e.g. Berry, 1971, 1975;	Berry, 1971
2. Entropy	$D = \sum_{i=1}^{n} s_i \ln(1/s_i)$, where s_i is the share of sales in segment *i*.	Patterns of firm revenues within portfolio; SIC hierarchical structure	Entropy calculated separately for 2- and 4-digit industries; difference in these scores is relatedness.	Diversification research; e.g. Palepu, 1985	Jacquemin and Berry, 1979
3. Wrigley-Rumelt categorizations	Categorization into one of nine categories based on three ratios: specialization, vertical, and related	Patterns of firm revenues within portfolio	Business is related if revenue from largest group of related activities (defined by researcher) is greater than 70 percent (related ratio) while no single industry's revenue is greater than 70 percent (specialization ratio)	Diversification research	Wrigley, 1970; Rumelt, 1974
4. Concentric	$D = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} s_i s_j r_{ij}$, where *s* is the percentage sales in industry *i* or *j*, and r_{ij} = 0 if *i* and *j* have the same three-digit code, 1 if they have identical two-digit codes (but not three-digit), and 2 if they have different two-digit codes	Patterns of firm revenues within portfolio; SIC hierarchical structure	Based on distances in the hierarchy of the SIC system; pairwise relatedness decreases as codes share only the same 3-digit, the same 2-digit, or different 2 digit codes, respectively	Diversification research; e.g. Montgomery and Wernerfelt, 1988	Caves, Porter and Spence, 1980

Table 2. Select Relatedness Measures

Measure	Mathematical Form	Empirical Base	Relatedness Component	Primary usage	Source
1. Scherer input-output matrix-based	$R = \cos\theta = \frac{\mathbf{x} \cdot \mathbf{y}}{\lVert\mathbf{x}\rVert \lVert\mathbf{y}\rVert}$, where cos θ is the Pearson correlation coefficient between industry categories **x** and **y**, which are centered vectors of technology inflows from all other industry categories	R&D flows based on patent usage data	Based on similarity between profiles of technology inflows	Tests of the resource-based view	Robins and Wiersema, 1995; Scherer, 1982
2. Occupational categories	$R_{ij} = \sum_{k \in K} (x_{i,k} - y_{j,k})^2$, where *x*, *y* are the normalized values of percent employees falling into occupational class *k* in industries *i* and *j*. These distances are further clustered into related industry groups (RIGs)	Occupa-tional classes	Based on similarity between occupational classes between industries	Tests of the resource-based view	Farjoun, 1990, 1994
3. Technological distance (patents)	$R_{ij} = \sum_{c} \Pr(i \mid c) N_{ic}$, where relatedness of firm *i* to industry *j* is a sum across patent classes *c* of the probability that patents of class *c* are assigned to industry *i* , multiplied by the number of firm patents in each class	Patents	Based on assignments made by the Canadian Patent Office of patents to industries of likely use, which in turn are matched to the US SIC system using Silverman's (1996) U.S. Patent Class—U.S. SIC concordance	Tests of the resource-based view	Silverman, 1996, 1999
4. Present Measure	$\tau_{ij} = \frac{J_{ij} - \mu_{ij}}{\sigma_{ij}}$, where *J* is the count of the number of firms diversifying into industries normalized using the hypergeometric distribution; τ is converted to a weighted distance matrix and shortest path scores through this matrix become inter-industry relatedness measures	All diversification moves in the US manufac-turing economy	Implicit in methodology and arising from economy of scope arguments	Tests of the resource-based view; examination of longitudinal expansion decisions	Present study

Figure 1. A sample network

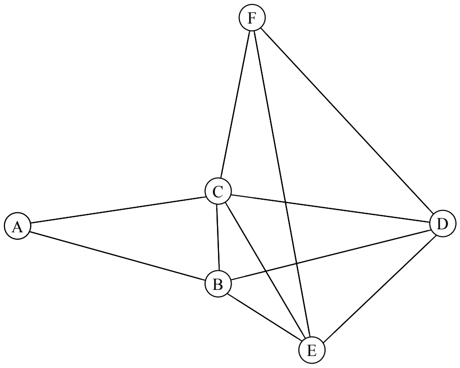

Figure 2. All interindustry relatedness scores: Four-digit SIC

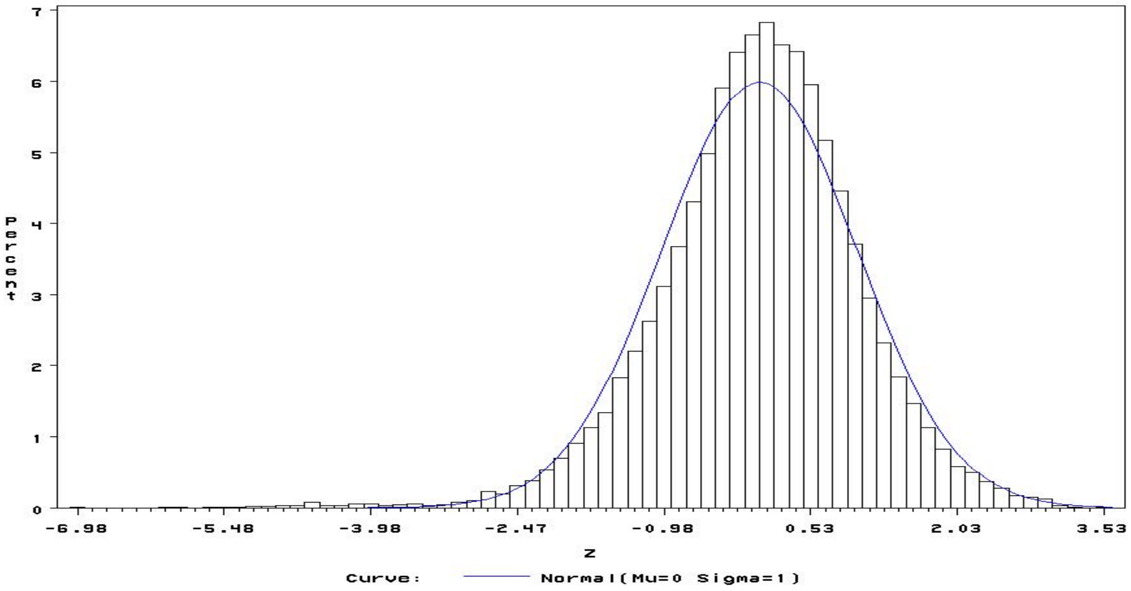

Table 3. Pearson correlation coefficients and descriptive statistics

	Variable	Mean	S.D.	1	2	3	4	5	6	7	8	9	10	11	12
1	Operations in two-digit industry in previous census	0.71	0.46	1											
2	Teece et al (1994) and Eq. (4)	12.2	8.96	*0.42	1										
3	General relatedness index	81.9	7.51	*0.30	*0.42	1									
4	Firm coherence	73.7	7.94	0.01	*0.12	*0.07	1								
5	Years of two-digit experience	18.1	11.81	*0.54	*0.32	*0.20	*0.15	1							
6	Ln (parent size)	12.9	1.8	*0.23	*0.18	*0.14	*0.14	*0.34	1						
7	Pre-entry industry growth rate	0.09	0.24	-0.02	-0.01	*0.07	0.03	0.01	0.04	1					
8	Four-firm concentration ratio	0.35	0.2	0.00	0.00	*-0.18	-0.04	0.01	*0.19	*-0.14	1				
9	Asset intensity	14.7	1.11	*-0.09	*0.16	*0.19	0.01	-0.02	-0.02	*0.07	*-0.08	1			
10	Average plant profitability in industry	0.16	0.32	-0.04	-0.04	-0.04	0.03	-0.02	0.01	-0.04	*0.08	0.00	1		
11	Proportion of startups that build vs. acquire	0.73	0.18	-0.02	*-0.17	*0.08	*-0.12	*-0.15	*-0.11	0.03	*-0.33	*-0.08	-0.02	1	
12	Industry R&D expense over net sales	0.02	0.03	-0.04	*-0.06	*-0.12	*-0.17	*-0.13	0.06	-0.01	*0.14	0.04	0.04	*0.10	1

**=p<0.01*

Table 4. Probit regression results for entry mode choice (build = 1)

Variable Description	(1)	(2)	(3)	(4)
Establishment				
Operations in two-digit industry in previous census		0.0677 0.0838		
Teece et al (1994) and Eq. (4)			0.0045 0.0038	
General relatedness index				0.0141** 0.0045
Firm				
Firm coherence		0.015*** 0.004	0.014*** 0.004	0.014*** 0.004
Years of two-digit experience		0.005 0.003	0.006* 0.003	0.005* 0.003
Ln (parent size)	-0.169*** 0.018	-0.196*** 0.020	-0197*** 0.020	-0.204*** 0.020
Industry				
Pre-entry industry growth	0.302* 0.133	0.304* 0.134	0.305* 0.134	0.291* 0.134
Four-firm concentration ratio	-0.291* 0.174	-0.211 0.176	-0.202 0.176	-0.140 0.178
Asset intensity (includes building and machinery)	0.075* 0.029	0.079** 0.029	0.072* 0.029	0.059* 0.029
Average plant 'profitability' in industry	-0.192* 0.108	-0.199* 0.111	-0.197* 0.111	-0.199* 0.112
The proportion of new (startup) firms that build vs. acquire	0.957*** 0.120	1.09*** 0.201	1.123*** 0.202	1.042*** 0.202
Industry R&D expense over net sales	1.718 1.285	2.785* 1.313	2.812* 1.314	3.217* 1.324
Intercept	0.390 0.530	-0.699 0.608	-0.582 0.606	-1.376* 0.649
-2logL (full model)	2199.41	2179.88	2179.17	2170.90

(*) = *p*<0.01, (**) = *p*<0.001, (***) = *p*<0.0001