MANAGEMENT SCIENCE
Vol. 55, No. 9, September 2009, pp. 1570–1585
ISSN 0025-1909 | EISSN 1526-5501 | 09 | 5509 | 1570

DOI 10.1287/mnsc.1090.1040

A General Interindustry Relatedness Index

David J. Bryce
Marriott School, Brigham Young University, Provo, Utah 84602, dbryce@byu.edu

Sidney G. Winter
The Wharton School, University of Pennsylvania, Philadelphia, Pennsylvania 19104;
and Australian School of Business, University of New South Wales, Sydney, NSW 2052, Australia,
winter@wharton.upenn.edu

For empirical work in the resource-based view of the firm, characterizing the resources that are responsible for firm growth is difficult because valuable resources are often tacit, ambiguous, or difficult to identify. This is a particular problem for empirical assessments that rely upon the concept of relatedness between resources to characterize the direction of growth of the firm. We tackle the problem for the general case by developing a general interindustry relatedness index. The index harnesses the relatedness information embedded in the multiproduct organization decisions of every diversified firm in the U.S. manufacturing economy. The index is general in that it can be used across industry contexts without requiring explicit identification of resources and it provides a percentile relatedness rank for every possible pair of four-digit Standard Industrial Classification 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.

Key words: relatedness; resource-based view; corporate strategy
History: Received January 10, 2007; accepted April 13, 2009, by Pankaj Ghemawat, business strategy. Published online in *Articles in Advance* July 10, 2009.

> 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, p. 86)

1. Introduction

According to resource-based theory, a firm's valuable and unique resources are at the root of its competitive advantage (Penrose 1959, Wernerfelt 1984, Barney 1991, Peteraf 1993, Conner and Prahalad 1996). However, identifying which of a firm's resources matter most for competitive advantage is no easy task. Although resource ambiguity may tend to protect competitive advantage, it presents difficult challenges for researchers in testing predictive theory. Consider the problem of predicting the direction of growth of the firm. Resource-based theory suggests that excess capacity in idiosyncratic resources, combined with externally determined opportunities, leads to expansion in directions related to a firm's existing resource stock (Penrose 1959). The predictive challenge arises in at least three ways: (1) identifying which resources are leveraged for growth; (2) determining how those valuable resources relate to competitive strength in potential target industries; and (3) developing some notion of how the firm chooses among the options.

Identifying the resources that are leveraged for growth is a challenge because the resources upon which competitive advantage rests are often bundled, tacit, intangible, or unobservable (Nelson and Winter 1982; Rumelt 1984; Peteraf 1993; Winter 1987, 1995). Second, determining how valuable resources relate to those effective in another industry is a challenge because the type of relatedness that should matter may be unclear. Relatedness is a multidimensional construct (Pehrsson 2006, Stimpert and Duhaime 1997), and different dimensions are likely to apply to different resources and in different contexts. Moreover, knowing which of the resources are the ones for which relatedness should matter in the target market is also problematic. Finally, even if the first two challenges can be resolved, the question of how to choose among viable target markets remains, but generally a firm will expand into those areas in which its resources deliver the greatest advantage (Penrose 1959).

In this paper, we respond to these challenges by developing a general interindustry relatedness index that can be applied across firm and industry contexts but that does not require explicit identification of resource type. Specifically, our approach employs the insight embodied in the *survivor principle* (Stigler 1968) by presuming that because existing firms are repositories for resources, skills, and knowledge, the activity patterns of going firms are good indicators of how resources and knowledge relate across diverse activities. To sidestep the problem posed by the difficulty of observing the actual resources that are leveraged for growth at the level of the firm, we posit that there is a characteristic basket of these resources for each industry. The question we answer is not what resources rest within any one industry basket in particular, but rather how the resources in a particular basket relate to the resources in other baskets. Knowing which idiosyncratic resources reside in a particular industry basket is not required for predictive success, because once it is known how that basket relates to every other industry basket, one knows in which candidate directions the leveraging of those unobservable idiosyncratic resources is likely to lead. The advantage of this approach is that it acknowledges that the characteristic resource baskets differ from industry to industry without requiring a specification of those differences. As we demonstrate in an illustrative application here, this permits empirical testing of hypotheses about relatedness without requiring the researcher to make a specific prior commitment as to the types of resources that are critical.

To identify the system of relationships among industries, the index harnesses the information embedded in the joint industry participation choices of every diversified firm in the U.S. manufacturing economy for the specific time period upon which the index is based. For this we specify the finest level of detail at which industrial "participation" can be effectively assessed, which we take to be the four-digit level of the Standard Industrial Classification (SIC) system. Our calculations yield a measured "distance" between the two industries in every pair of four-digit industries in the U.S. manufacturing sector, where low distance corresponds to high relatedness. Our methods 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.

The index is applicable to a wide range of problems in strategic management, corporate finance, and economics because it provides a plausible measure of the relative strength of association between every pair of manufacturing industries. The index may be particularly applicable to empirical examinations of strategic 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 et al. 2000; Teece et al. 1997; Grant 1996; Kogut and Zander 1992; Helfat and Eisenhardt 2004), because these perspectives typically require assessment of the degree of overlap, knowledge, or relatedness between one firm activity and another. Similarly, concepts of relatedness are fundamental to discussions of how firms search for new market-entry opportunities that economize on existing resources as they build new capabilities (Bryce 2003; Coff 1999; Silverman 1999; Teece 1980, 1982); how capabilities develop from sequences of decisions that are made in the context of resources in hand (Helfat and Raubitschek 2000, Helfat and Lieberman 2002, Helfat and Eisenhardt 2004); or how the ability to share firm-specific resources across activities results in higher levels of firm performance (Teece 1982, Peteraf 1993, Mahoney and Pandian 1992, Teece et al. 1997). Applications of the measure to the study of longitudinal patterns of diversification and firm growth are especially promising because the measure allows sequential analysis of the introduction of new industries into a firm's portfolio, one activity at a time.

This paper proceeds as follows. In §2, we provide a brief review of how concepts of relatedness and diversification have been used in the literature, supply the theoretical rationale for our particular approach, and propose solutions to some methodological problems that arise in measuring relatedness using a survivor-based approach. In §3, we develop the index, and in §4, we offer a test of predictive validity. Section 5 concludes with a discussion of potential applications.

2. Background and Theory

2.1. Measures of Relatedness

Measures of relatedness are designed to assess the degree of commonality (of some sort) within pairs of activities. They differ in their logic and uses from diversification measures, which are typically designed to support evaluation of a diversification strategy at the firm or portfolio level. Whereas diversification measures capture the state of a corporate portfolio at a point in time, relatedness measures can be used to characterize the flow or transition from state to state. Diversification measures do rest on underlying relatedness assessments, however. In standard diversification measures, relatedness is typically computed based on hierarchical distance within the SIC structure—a course that implicitly relies on the designers of the SIC system to have already answered the basic question.[1]

[1] Most activity-level relatedness measures can be aggregated into a composite firm-level measure of diversification by using some weighting scheme across activities (e.g., Robins and Wiersema 1995). The common diversification measures found in the literature,

Relatedness components in standard diversification measures cannot effectively serve as stand-alone relatedness indicators because the hierarchical structure of the SIC system does not represent an underlying relatedness scale. 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. This virtually guarantees that the knowledge about how to produce a functionally similar product lies scattered around the SIC system. For some two-digit SIC 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) supplies 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).[2]

Perhaps most importantly, the SIC hierarchy does not consistently reflect relationships among valuable resources in the ways that firms actually combine them to create value. Our approach of inferring relatedness from the aggregate activity combinations of firms provides a strong resource-based measure of relatedness between industries because it reflects the unobservable ways that firms share resources among industry activities. A measure rooted in actual resource combinations inside firms has the advantage of reliably merging both demand-side and supply-side considerations in the way firms deploy resources. Peteraf and Bergen (2003), for example, note that resource substitution effects by rivals are an important potential source of erosion of a firm's resource-based competitive advantage. Rivals who aim to compete on the basis of similarity of resource use, not just type, may go unnoticed by firms that regard as competitors only those firms that have similar resources and produce similar products (demand side). The methodology used here is advantageous in this regard because it also captures use relationships (supply side) while avoiding dependence on the relatively arbitrary industrial taxonomy of the SIC.

There are only a few relatedness measures with a non-SIC-hierarchy-based foundation (e.g., Robins and Wiersema 1995, Silverman 1999, Farjoun 1994, Coff 1999). These are activity-to-activity constructs that have been developed by researchers to test propositions suggested by the resource-based view of the firm. The same is true of the measure developed here. However, this paper focuses on the development of the relatedness measure itself. Unlike the extant relatedness measures, it is developed as a general index with a broad range of potential applications rather than to accomplish a specific empirical task. It also has the advantage that it does not presume prior identification of a key resource class such as human resources (Chang 1992, Farjoun 1994, Coff 1999), patents (Silverman 1999), or technology flows (Robins and Wiersema 1995)—the relevance of which may vary with activity—prior to computing relatedness scores. By moving away from a focus on single resource categories, our measure seeks to capture the aggregate patterns of shared know-how or capabilities (Teece 1982) that are at the root of economies of scope and resource combination decisions.

2.2. Theoretical Rationale

Ultimately, the test of the validity of our index will be its predictive power in a variety of significant empirical settings, going well beyond the illustrative test we present later in this paper. At this stage, we can only support the index by setting forth the assumptions and arguments that motivated our approach to its construction. Such an account may provide some theoretical guidance regarding the appropriate use of the index, and it may also suggest how it might be improved or how the empirical validity of its underlying logic might be checked more directly.

We adopt, first, the premise that the resource-based view (Peteraf 1993, Barney 1991, Wernerfelt 1984) is substantially correct in its assessment of the forces affecting the directions of firm growth. We assume that patterns of corporate diversification and expansion are shaped in a fundamental and sustained way by logic of economic efficiency (Teece 1980). 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 (Teece 1982). 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 (e.g., Teece

such as the entropy measure (Jacquemin and Berry 1979, Palepu 1985, Hoskisson et al. 1993), the Herfindahl-based measure (Berry 1971, Gollop and Monihan 1991), or the concentric index (Caves et al. 1980, Montgomery and Hariharan 1991, Montgomery and Wernerfelt 1988), have this sort of grounding in a relatedness measure and typically, but not always (e.g., Gollop and Monihan 1991, Berry 1971), rely upon SIC hierarchy-based relatedness.

[2] A possible alternative to SIC hierarchy-based relatedness is the categorical method of relatedness identification based on researcher judgment (e.g., Wrigley 1970, Rumelt 1974). However, these methods apply a portfolio-level, not activity-level, category designation. These methods are 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).

1980, Panzar and Willig 1981, Lemelin 1982). Scope economies can arise in a short-run context because indivisibilities or other considerations have led the firm to commit to an array of tangible resources that is underutilized by the existing product mix (Penrose 1959, Teece 1982). In the long run, however, it seems likely that intangible resources, especially specialized types of knowledge, provide the most fundamental and durable source of scope economies. 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 indefinitely large extent. By virtue of the nonrivalrous character of information, there is no limit to the application of specialized knowledge that is intrinsic to its own nature. Although there are costs to replicating knowledge, these costs operate more strongly on the pace at which leveraging takes place than on its ultimate extent. The latter is shaped primarily by the "demand side"; it is the environment that determines the size of the domain in which profitable application of given knowledge can ultimately take place.

At any given time, the patterns of corporate participation in different industries reflect the cumulative effect of the operation of this knowledge-based efficiency logic in the past—along, of course, with whatever other causal determinants and random effects may be involved. Thus, current diversification patterns reflect a series of choices in which firms leveraged existing knowledge and also acquired or developed new knowledge complementary to existing knowledge. Although other determinants may well have shaped some particular choices, we would expect that, in the large and in the long run, firm scope tends to reflect the underlying knowledge structure. A firm that consistently acted in defiance of that structure when choosing new activities would repeatedly face liabilities of inexperience. In this sense, our argument may be viewed as relying upon the *survivor principle* in that it presumes that what firms actually do makes economic sense.[3] 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 economywide implication of such firm-level sense is what our index seeks to capture. In relying on this principle, we do not presume that it operates with great promptness or precision, though we would argue that it is probably stronger for firm scope than for size. 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 are weak—but this is not so likely when they are strong.

Because 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 (1991, p. 230) derived from Lenin—"Why is the economy not run as one big factory?"—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.

Although we argue that the strongest of the economic forces is likely to arise from economy of scope and resource relatedness considerations, it is important to consider alternative views of the forces underlying patterns of diversification. Consider market power and agency theory explanations. Market power explanations suggest that diversified firms take positions in multiple markets to practice anticompetitive behavior by engaging in collusive practices, or they use multimarket competition to stifle competition and drive up profits (Montgomery 1994, Bernheim and Whinston 1990). Agency theory suggests that when monitoring by owners is difficult, managers spend free cash flow in empire-building activities that further their personal interests to the likely overall detriment of firm owners (Jensen 1986). Although these are important perspectives, neither market power nor agency arguments offer any general theoretical guidance on where one should expect firms to make investments. They are theories of *why* firms diversify and not theories of *where* firms diversify or how the pattern of diversification unfolds. Thus, these theories are silent on the question of how relatedness affects the directions of diversification, if at all. Although there can be little doubt that such motives explain

[3] As originally stated by Stigler (1968, p. 73) for the context of firm size, the survivor principle is that "the competition of different sizes of firms sifts out the more efficient enterprises."

many instances of diversification, the motives do not imply that the resulting moves are other than randomly distributed among industries. This is equally true for imitative motivations to diversify. As a result, the sample frequency of any one particular industry pair occurring inside firms is not likely to be significantly influenced by agency, market power, or imitative considerations. Certainly, these motivations could not persist for long within particular industry pairs in the absence of economic logic. Nevertheless, although it is doubtful that expansion choices rooted in such motivations are systematic enough to influence the index, if they are, then relatively strong economic forces are probably in play and our methodology will capture them.

Stimpert and Duhaime (1997) and Pehrsson (2006) suggest that managerial logic for diversification is multidimensional and that there are logics for diversifying, such as commodity or financial relatedness, that are not captured effectively by standard diversification indices. Nevertheless, unless these motivators are underscored by strong economic forces, they are unlikely to be systematic across large numbers of joint industry participation decisions. If managerial conceptualizations of relatedness are systematically employed by managers in actual diversification moves (e.g., Prahalad and Bettis 1986, Grant 1988), our measure will reflect them.

In any single point of time, the frequencies of particular pairs of industries may be, in a sense, out of equilibrium, reflecting firm experiments or even fads that may not be driven by durable considerations. However, the likelihood of observing any particular industry pair combination, whether an "experiment" or otherwise, rises in the length of time that it endures inside firm portfolios, which itself increases in the strength of the economic logic on which the experiment is based. Thus, the frequency of such experiments in the sample should be proportional to the strength of the economic force that governs the experiment, and this works in favor of the index. Purely random experiments are expected to be relatively few and randomly distributed in the sample so that they are unlikely to systematically bias the index.

2.3. Issues in Developing a General Index

Basing a measure of relatedness on actual diversification patterns raises several important methodological issues 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. As we argued above, choices about expansion will be predominantly driven by efficiency and knowledge-use logic, but some combinations may result from managerial experimentation, agency, imitation, or other reasons. The key is to capture what is systematic in the overall patterns. In any given industry pair, the number of "accidents" will be greater 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. It is also possible that richer competition in active industries could weed out such accidents more quickly. These issues are addressed by noting that the key to harnessing the information content in diversification moves 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. The "coherence" methodology of Teece et al. (1994) supplies a normalization approach to resolve this issue. Teece et al. (1994) count the frequency of pairs of SIC industries appearing jointly in firm portfolios and normalize this frequency to identify cases in which pairs of industries are appearing more frequently than randomness would suggest. The Teece et al. (1994) approach is the effective starting place for development of the general index.

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 weakly to other activities in the portfolio. This second issue is addressed by weighting the normalized dyad 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. If activities can be effectively combined through market mechanisms (Teece 1980), there may be no need to combine them within the firm. However, the balance of costs and benefits may change over time as some firms gradually extend the scope of their capabilities. This could bring activities inside the firm that were previously contracted. To accommodate this idea, our measure includes a provision that fills in the relatedness picture in cases where the direct evidence of actual joint participation is entirely absent.

To motivate our approach to filling in the relatedness picture when two industries in a pair never appear inside a firm portfolio, consider the problem of determining the driving distance between two cities, A and C, that are not connected by any roads on a map. Upon examination, one finds that city A is connected by road to city B, which in turn is connected by road to city C. Therefore, the actual driving distance between A and C is simply the sum of the distance A to B and B to C. If one is driving, no other distance can be meaningfully considered because existing roads must be followed. To implement this simple idea, we create a network representation of the weighted relatedness distances between industry nodes and compute the shortest path scores between nodes.[4] This procedure produces relatedness scores based on proximity in the network for activities that are not combined in any firm, and it replaces direct distances with shortest path scores when the shortest path distance is less. Replacing direct distances with shortest path scores is like finding a shortcut between two cities by driving through a third city.

The shortest path methodology has the virtue of capturing the knowledge structure among industries in a way that is durable, even if underlying industry combination decisions change on the margin from year to year. By replacing direct distances with shortest path distances, the methodology effectively finds the most related linkages between industries in the network and eliminates other, perhaps more fleeting, linkages. The net effect is that the knowledge structure identified by the shortest path network is expected to be stable, and this ultimately facilitates the index's use for predictive purposes at points in time different from the measurement year.[5]

3. Construction of the Index

3.1. Data

Our data are 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 data also has the distinct advantage of supplying a census rather than a sample of firms; operating data on all multiunit firms that appeared in the 1987 Census of Manufactures (SIC 2000–3999) is included.[6]

3.2. 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 industries i and j is $J_{ij} = \sum_k C_{ik} C_{jk}$. As explained above, raw counts of the number of firms operating in each industry dyad 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 if firms were assigned to industries at random (cf. Teece et al. 1994).

To accomplish this adjustment, the distribution of J_{ij} must be derived. For now, call this random variable X_{ij}.[7] Our task is to compute the probability that x out of K firms receive a random assignment to both

[4] 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, 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 $p \in P = v_1 e_1 v_2 e_2 \cdots v_{k-1} e_{k-1} v_k$, $v_i \in V$, $e_i \in E$ is the sum of the costs of the 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.

[5] When a direct distance is replaced with an indirect distance, it implies that the absence or weakness of joint participation in the dyad is not indicative of the true relatedness. It is not hard to specify circumstances in which this might be the case; however, contrary circumstances can also be imagined. As noted below, the replacements assure that the triangle inequality holds throughout for our distance measures, so that the notion of "distance" is meaningful. We consider this to be a compelling advantage of making the substitutions.

[6] Here we define a firm as multiunit 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.

[7] Teece et al. (1994) identify the distribution, but they do not derive it in their paper. We found it necessary to derive the distribution to

industry i and industry j. For this random model, we take the industry sizes n_i and n_j and the population size K to be given numbers, but postulate the equal likelihood of all distinct membership rosters for these industries that can be formed from a given population of K multi-industry firms. Some of these rosters for i and j overlap to the extent x. The question is how many of these there are. First consider the ways of specifying those firms that sit in the x positions of the overlap from among the firms active in industry i. This is equivalent to the number of ways of selecting x from a total of n_i firms, or $\binom{n_i}{x}$. With the firms in the overlaps specified, there are $(n_j - x)$ positions in the n_j roster to be filled with firms that are not also active in i. The number of ways of filling these is 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 distinct ways of choosing a roster for industry j that is consistent with the specified overlap is the product of the answer to the first question and the answer to the second, or $\binom{n_i}{x}\binom{K-n_i}{n_j-x}$.[8] To turn this count into a probability for x, we divide it by the number of possible ways of specifying the membership of j in total; i.e., when the constraint of the overlap is dropped, which is $\binom{K}{n_j}$. Thus, the level of randomly occurring joint participation 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,

$$P[X_{ij} = x] = \frac{\binom{n_i}{x}\binom{K-n_i}{n_j-x}}{\binom{K}{n_j}}. \tag{1}$$

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

$$\sigma_{ij}^2 = \mu_{ij}\left(1 - \frac{n_i}{K}\right)\left(\frac{K-n_j}{K-1}\right).^{9} \tag{3}$$

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. We note that the existence of pairs that are represented more frequently than suggested by the random model necessarily implies a complementary set of relatively underrepresented pairs. Underrepresentation does not imply some sort of negative relatedness, but only that 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. Because Equation (4) is based on raw industry participation counts, it is a coarse measure of the extent to which activity combination *ij* is economically important. 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. In a broadly diversified firm, two activities each delivering only 1%–2% 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. As argued above, 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. 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

check what turned out to be minor typos in the original publication. Because doing so clarifies the setup of the problem, we include the brief exposition here. The original article is reprinted, with most if not all of the errors corrected, in Langlois et al. (2003).

[8] Because 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.

[9] 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 n_i/K. Because there are n_j firms in K, each with probability n_i/K of being assigned to industry i, the expected number of firms assigned to both industry i and industry j is $n_j(n_i/K)$. For further information on the hypergeometric distribution, see Feller (1957).

averaged across all firms operating in the dyad to create the dyad weight. The average weight S_{ij} produced by all firms operating in the dyad is

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

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, SICs 2386, 2371, 3263.[10] 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. As noted above, this issue is addressed by solving for the shortest path distance between every pair of nodes in the weighted distance matrix. The method 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).

[10] These codes are Leather and Sheep-lined Clothing, Fur Goods, and Fine Earthenware (Whiteware) Table and Kitchen Articles, respectively.

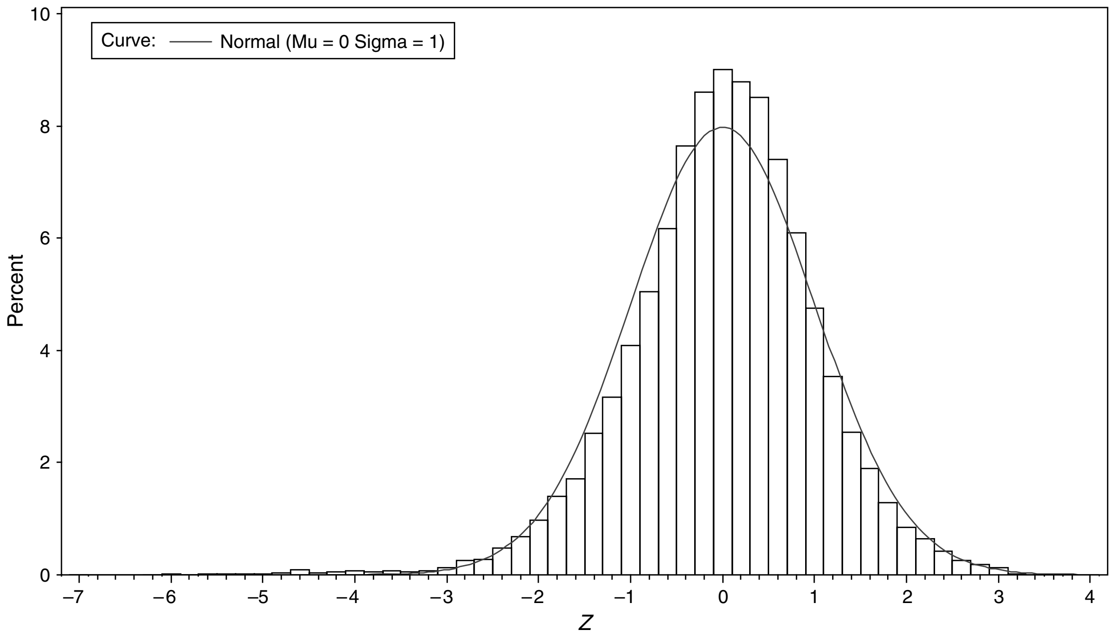
To complete construction of the index, the weighted distance matrix, which is now filled with shortest path scores, is converted to a similarities matrix, where the greatest values rather than the lowest values represent the highest relatedness. This is done simply by subtracting each computed path length 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 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. Second, 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% of industry dyads are less related than the focal score, whereas 30% are more related. Plots of the distribution of all normalized (not percentile) dyad relatedness index scores are shown in Figure 1. Note that Figure 1 represents only the distribution of dyadic relatedness scores; it is not indicative of the extent of relatedness within firm portfolios.

3.3. Examples of Index Scores

A few examples of index scores illustrate the ability of the index to capture relatedness relationships among industries and 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.[11] The relatedness here squares with what intuition might suggest; the advantage of the index is that it provides a quantitative basis for comparison to other dyads. The two most unrelated industries are SIC 2097, Ice, and

[11] 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.

Figure 1 All Interindustry Relatedness Scores: Four-Digit SIC



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.

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. SIC hierarchy-based relatedness methods typically consider industries that are differently classified at this high level 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. Indeed, such activities appear to be vertically related.

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. Although 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. (PRNewswire 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 nonobvious 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 the 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 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.

4. Test of Predictive Validity

The predictive value of the 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 to predict investment decisions by firms under hypotheses that those decisions are rooted in resource-based logic. (Needless to say, there is no example of quantitative prediction in the domain of science that does not rely on an assumption that something measured at one time is still holding that same value at a later time.)

4.1. Entry Mode Choice

To test the predictive validity of the measure (understood as the degree to which a measure of a concept shows the expected statistical relationship with some recognized outcome; Lubatkin et al. 1993, p. 436), we employ a more conservative test than is represented by examining the direction of corporate growth directly. If the index represents relationships between industries in valuable, idiosyncratic resources, we argue that it should predict the *mode* of entry of an expanding firm. The test is a demanding one in that it asks whether the information in the index can predict the choice 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. 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.

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 showed a positive correlation between the relatedness of existing activities and the target industry (Yip 1982). 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 new production function in the target industry. The requisite coordinating information for productive activity is partly imported into a new plant 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 plant itself, such as in its structures, layout, or machines. A firm holding very specific and highly technical knowledge 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).

By contrast, some firms may have resources that would be helpful in a target industry yet lack the specific knowledge required to create the necessary production functions themselves, and therefore resort to acquisition for entry. Acquisition may be the only option when the firm lacks the specific knowledge that would make it an effective builder.[12] A functioning plant 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. In such cases, less supporting knowledge is required at the firm level to ensure continued operation in the plant. In most cases, an acquisition of a going plant also includes acquisition of the capabilities embedded in the plant, including the tacit knowledge residing with personnel.

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 plant.*

[12] Some plants may be built on behalf of the focal firm by specialist engineering firms who bring technical knowledge to get a "turn-key" plant up and running (e.g., Arora et al. 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, because 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.

4.2. Data and Methods

The sample for the analysis includes all plants 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 plants is 4,721. However, because of missing values for select covariates (e.g., industry research and development (R&D) expenditures), the number of plants included in the regression analysis is reduced to 1,706.[13] The choice of entry mode is modeled as a dichotomous variable, where 1 is entry by build and 0 is entry by acquisition, utilizing a probit specification. Regressions model the likelihood that a firm chooses to build (versus acquire) its way into a new industry.[14] 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.

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 plants 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 in standard diversification measures is able to distinguish entry mode 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. The most-related methodology codes the relatedness to the target for a firm based on the closest existing activity in the portfolio. In the two-digit case, relatedness is coded 1 if any existing activities share the same-two digit class with the target. In the case of the Teece et al. (1994) measure and the general index, relatedness to the target is first computed for all activities and then the most related score is utilized. The logic is that the firm leverages the knowledge contained in its closest activity as it makes a build or acquire choice. A positive sign is expected on relatedness coefficients, indicating that relatedness increases the probability of a build choice.

Coherence. Firm 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, inclusion of this variable provides an important control on the way that past portfolio choices influence entry mode.

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 because 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) are included (see the appendix for a detailed description). Also included are controls for *industry build propensity* and *R&D intensity*. The latter two variables are particularly important as further controls on the strategic effects of specific knowledge. R&D-intensive industries are likely to require the development of specialized resources for effective competition, and holders of specialized resources are more likely to enter by building. The *build propensity* variable (implemented at the industry level) is a control that reflects the extent to which entrant firms

[13] R&D intensity is calculated at the industry level based on COMPUSTAT (see the 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 those runs. We do not include those results here.

[14] We have also analyzed whether the general index predicts the choice of industry, a test we regard as less stringent than the one shown here. The analysis shows a strong tendency across industries for building firms to enter industries that are, statistically, significantly more related to entering firms' activities than are those industries to the activities of firms that choose to build elsewhere. In other words, industries tend to attract builders that are statistically the most related to them among all potential builders.

Table 1 Pearson Correlation Coefficients and Descriptive Statistics

Variable	Mean	SD	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 Equation* (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 *Preentry 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 start-ups 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 divided by 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

*$p < 0.01$.

naturally build in some industries and not in others. This may reflect the specific knowledge required in an industry but it also may reflect low start-up costs, where entrant firms nearly always build rather than acquire plants, even when acquirable plants are available. These additional controls are interesting in their own right, and their inclusion establishes a more demanding test of the predictive power of the relatedness variables. Pearson correlation coefficients for all variables are shown in Table 1.

4.3. Results

Results of the probit regression analyses are shown in Table 2. Strong support is found for the general index as a predictor of entry mode choice, which is highly significant at $p < 0.001$. Model 4, which incorporates the general relatedness index, performed the best overall.

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 weighting and shortest path search contributes important information to the task of assessing relatedness. The general index differentiates between build and acquire entry modes, whereas the other measures do not.

All firm-level variables, *firm coherence*, *number of years since earliest two-digit experience*, and other controls such as *preentry industry growth*, and the natural log of *parent size*, are significant in the expected direction. The negative coefficient on firm size implies, as expected, that firms with greater size are likely to have greater resources to deploy toward acquisition. Industry controls of concentration, asset intensity, and profitability turn out generally as expected, although concentration is only significant in Model 1. The sign on average plant profitability in the target industry is negative. This is consistent with the idea that firms seek to enter profitable industries by acquiring profitable plants. It also supports the idea that firms from afar—those without detailed productive knowledge—will enter, but that they will be restricted on average to entering by acquisition rather than by build.

The build propensity variable is positive and highly significant. To the extent that such propensities are related to knowledge structures, the fact that the general index demonstrates residual strength in the presence of the already strong effects for this variable is noteworthy.

R&D intensity was positive and significant as expected. This suggests that R&D intensity contributes to the explanation for selecting build versus acquire over and above the knowledge specificity captured by our relatedness measures and other theoretical variables. This is an industry-level variable and not one that is tied specifically to the firm. Although the general index purports to capture specific knowledge, it does so in these regressions at the firm level only, and this apparently leaves some residual influence at the industry level to be captured by the R&D intensity variable. The significance of the variable offers support for our theoretical arguments that specific knowledge leads firms to build because they are uniquely qualified to supervise the creation of a production function that is operative in a new plant.

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. More importantly, the index demonstrates the ability to significantly measure and predict firm behavior in cases in which theorizing would suggest highly specific knowledge is in use, and it does so in the presence of other firm- and industry-level measures that might otherwise be expected to capture those effects.

Table 2 Probit Regression Results for Entry Mode Choice (Build = 1)

Variable description	(1)	(2)	(3)	(4)
Plant				
Operations in two-digit industry in previous census		0.0677		
		0.0838		
Teece et al. (1994) *and Equation* (4)			0.0045	
			0.0038	
General relatedness index				0.0141**
				0.0045
Firm				
Firm coherence		0.015***	0.014***	0.014***
		0.004	0.004	0.004
Years of two-digit experience		0.005	0.006*	0.005*
		0.003	0.003	0.003
ln(*parent size*)	−0.169***	−0.196***	−0.197***	−0.204***
	0.018	0.020	0.020	0.020
Industry				
Preentry industry growth	0.302*	0.304*	0.305*	0.291*
	0.133	0.134	0.134	0.134
Four-firm concentration ratio	−0.291*	−0.211	−0.202	−0.140
	0.174	0.176	0.176	0.178
Asset intensity (includes building and machinery)	0.075*	0.079**	0.072*	0.059*
	0.029	0.029	0.029	0.029
Average plant "profitability" in industry	−0.192*	−0.199*	−0.197*	−0.199*
	0.108	0.111	0.111	0.112
Industry build propensity	0.957***	1.090***	1.123***	1.042***
	0.120	0.201	0.202	0.202
R&D intensity	1.718	2.785*	2.812*	3.217*
	1.285	1.313	1.314	1.324
Intercept	0.390	−0.699	−0.582	−1.376*
	0.530	0.608	0.606	0.649
$-2 \log L$ (full model)	2,199.41	2,179.88	2,179.17	2,170.90

* $p < 0.01$; ** $p < 0.001$; *** $p < 0.0001$; $n = 1{,}706$.

5. Discussion

Given the findings here, the general relatedness index is expected to be a useful tool for assessing interindustry relatedness in virtually any context requiring such a measure. We argue that it essentially captures the knowledge relatedness structure underlying the U.S. manufacturing economy in the ways that firms actually combine resources to create value. Along with the methodology of statistical normalization, value-added weighting and averaging, and shortest-path substitution, the knowledge structure relationships identified 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 each time it is used. Its strong empirical base—all diversified firms in the U.S. manufacturing economy—makes repeated construction costly and difficult. 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. The methodology can also be used on the North American Industrial Classification System.

The measure is not without limitations. It is currently computed only for manufacturing industries. Future index development efforts might focus more broadly on the relationships between all industries in the economy using the same methodology. Also, the index will not be particularly useful for studies in which the researcher has hypothesized that a particular resource type drives expansion decisions, or to distinguish specific types of relatedness. For such studies, researchers will no doubt continue to create relatedness constructs for specific purposes.

The potential applications of the index are many, but the index holds particular promise in studies of firm expansion and diversification, where it offers new empirical tools to test theoretical logic based on the resource-based view of the firm. Outlined below are three specific applications for which the general index promises to be particularly useful.

5.1. 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 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. Because sequential choices about market entry are strongly influenced by and ultimately shape the capability profile of the firm, such choices are expected to have significant influence on firm performance over time. The empirical opportunity is to use the general index to plot sequential market-entry choices with respect to the relatedness between new activities and existing ones, and measure the influence of such relatedness patterns on intertemporal firm performance.

5.2. Related vs. Unrelated Diversification

Several alternative approaches exist for converting pairwise relatedness scores to a resource-based measure of diversification. One simple approach is to take a straight average of interindustry relatedness scores inside a portfolio, or to weight these averages by size of business in employees, sales, or some other indicator. Another approach is to embed the index scores into a Herfindahl calculation, similar to the approach taken by Gollop and Monihan (1991). Following the Teece et al. (1994) coherence indicator, yet another approach is to compute the minimum spanning tree of portfolio activities and average the scores on the tree, as we did when we created a control variable in the exercise above.

One potential application of a diversification measure based on the index is to distinguish efficiency- and resource-based motivations for expansion from agency-based motivations. Montgomery (1994, p. 174) wished for such a test: "Looking ahead, it would be very useful to have empirical tests that would help us discriminate between and evaluate the relative importance of the resource-base and agency theory views of diversification. Devising such a test may await a deeper understanding of the resources that can be beneficially leveraged across markets, and the critical differences between deploying these in a firm or market setting."[15] The empirical challenge is to determine whether variables that might proxy agency problems tend to be effective in explaining the incidence of cases where firms expand into industries that are too far removed from the firm's existing resource base to make the move explicable on relatedness grounds.

With respect to portfolio structure, it is also possible, using the index, to construct detailed profiles of firm portfolios and fine-grained measures of relative relatedness among all industrial activities in each portfolio. Examination of intraportfolio 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.

5.3. Capability Dynamics

Helfat and colleagues (Helfat and Raubitschek 2000, Helfat and Peteraf 2003, Helfat and Eisenhardt 2004) have drawn particular attention to the dynamics of knowledge and capabilities and the implication of these dynamics for the changing scope of the firm and the evolution of firm capabilities. Firms draw upon knowledge systems to create particular product sequences (Helfat and Raubitschek 2000); they leverage intertemporal economies of scope in incremental business entry and exit (Helfat and Eisenhardt 2004); and their choices lead to capability evolution (Helfat and Peteraf 2003). Together, these studies are fundamentally about resource-based change and how knowledge guides investment opportunities and influences capabilities. The general index holds strong promise for helping to push the boundaries of understanding of these phenomena because it can be used as a tool in several specific ways: (1) for measuring the degree of incremental change; (2) for sorting those changes on relative relatedness or knowledge specificity dimensions; (3) for tracing out changes in capabilities through time; or (4) for determining how developmental and evolutionary pathways influence firm performance.

This is only a sampling of the possibilities that are available. Hopefully, with the help of the general index, researchers can address with renewed vigor a broad range of important empirical problems of the sort outlined here.

Acknowledgments

The authors thank Jeff Dyer, Connie Helfat, Robert Hoskisson, Steven Klepper, Anne Marie Knott, Dan Levinthal, Phanish Puranam, Harbir Singh, Louis Thomas, anonymous referees, and participants at the BYU-Utah Winter Strategy Conference for helpful comments and suggestions. This research was conducted at the Center for Economic Studies, Census Bureau. 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 ensure that no confidential Census data are revealed.

[15] We thank an anonymous referee for reminding us of Montgomery's (1994) statement.

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 because large firms are more likely to acquire given greater access to external financial resources (Chatterjee and Singh 1999).

Preentry 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 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.

Asset intensity measures the capital requirements for entrants. It is calculated as the natural log of industry investments in plant and 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 sort that require specific knowledge—such as in highly technical industries requiring heavy expenditures in R&D. Thus, rationale for the sign of either direction can be developed and we make no prediction about the sign of this variable.

Average plant 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 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 searches 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 (start-up) firms that build versus 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. We expect a positive sign.

R&D intensity is the extent to which 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 plants 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.

References

Arora, A., A. Fosfuri, A. Gambardella. 2001. Specialized technology suppliers, international spillovers and investment: Evidence from the chemical industry. *J. Development Econom.* **65**(1) 31–55.

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

Bernheim, D. B., M. D. Whinston. 1990. Multimarket contact and collusive behavior. *RAND J. Econom.* **21**(1) 1–26.

Berry, C. H. 1971. Corporate growth and diversification. *J. Law Econom.* **14**(2) 371–383.

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

Caves, R. E., M. Porter, A. M. Spence. 1980. *Competition in the Open Economy: A Model Applied to Canada*. Harvard University Press, Cambridge, MA.

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, Philadelphia.

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

Chatterjee, S., J. Singh. 1999. Are tradeoffs inherent in diversification moves? A simultaneous model for type of diversification and mode of expansion decisions. *Management Sci.* **45**(1) 25–41.

Coase, R. H. 1937. The nature of the firm. *Economica* **4**(16) 386–405.

Coase, R. H. 1991. Nobel lecture: The institutional structure of production. Reprinted in O. E. Williamson, S. G. Winter, eds. *The Nature of the Firm*. Oxford University Press, New York, 227–235.

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

Conner, K. R., C. K. Prahalad. 1996. A resource-based theory of the firm: Knowledge versus opportunism. *Organ. Sci.* **7**(5) 477–501.

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

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

Farjoun, M. 1994. Beyond industry boundaries: Human expertise, diversification and resource-related industry groups. *Organ. Sci.* **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, UK.

Gollop, F. M., J. L. Monihan. 1991. A generalized index of diversification: Trends in U.S. manufacturing. *Rev. Econom. Statist.* **73**(2) 318–330.

Grant, R. M. 1988. On "dominant logic," relatedness and the link between diversity and performance. *Strategic Management J.* **9**(6) 639–642.

Grant, R. M. 1996. Toward a knowledge-based theory of the firm. *Strategic Management J.* **17**(Winter) 109–122.

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

Helfat, C. E., K. Eisenhardt. 2004. Inter-temporal economies of scope, organizational modularity, and the dynamics of diversification. *Strategic Management J.* **25**(13) 1217–1232.

Helfat, C. E., M. J. Lieberman. 2002. The birth of capabilities: Market entry and the importance of pre-history. *Indust. Corporate Change* **11**(4) 725–760.

Helfat, C. E., M. Peteraf. 2003. The dynamic resource-based view: Capability life-cycles. *Strategic Management J.* **24**(10) 997–1010.

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

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

Jacquemin, A. P., C. H. Berry. 1979. Entropy measure of diversification and corporate growth. *J. Indust. Econom.* **27**(4) 359–369.

Jensen, M. C. 1986. Agency costs of free cash flow, corporate finance, and takeovers. *Amer. Econom. Rev.* **76**(2) 323–29.

Kogut, B., U. Zander. 1992. Knowledge of the firm, combinative capabilities, and the replication of technology. *Organ. Sci.* **3**(3) 383–397.

Langlois, R. N., T. F.-L. Yu, P. L. Roberston, eds. 2003. *Alternative Theories of the Firm*. Edward Elgar, Cheltenham, UK.

Lemelin, A. 1982. Relatedness in patterns of interindustry diversification. *Rev. Econom. Statist.* **64**(4) 646–657.

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

Mahoney, J. T., J. R. Pandian. 1992. The resource-based view within the conversation of strategic management. *Strategic Management J.* **13**(5) 363–380.

Montgomery, C. A. 1994. Corporate diversification. *J. Econom. Perspectives* **8**(3) 163–178.

Montgomery, C. A., S. Hariharan. 1991. Diversified expansion by large established firms. *J. Econom. Behav. Organ.* **15**(1) 71–89.

Montgomery, C. A., B. Wernerfelt. 1988. Diversification, Ricardian rents, and Tobin's q. *RAND J. Econom.* **19**(4) 623–632.

Nelson, R., S. Winter. 1982. *An Evolutionary Theory of Economic Change*. Harvard University Press, Cambridge, MA.

Palepu, K. 1985. Diversification strategy, profit performance and the entropy measure. *Strategic Management J.* **6**(3) 239–255.

Panzar, J., R. Willig. 1981. Economies of scope. *Amer. Econom. Rev.* **71**(2) 268–272.

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

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

Peteraf, M. A. 1993. The cornerstones of competitive advantage: A resource-based view. *Strategic Management J.* **14**(3) 179–191.

Peteraf, M., M. E. Bergen. 2003. Scanning dynamic competitive landscapes: A market-based and resource-based framework. *Strategic Management J.* **24**(10) 1027–1041.

Prahalad, C. K., R. A. Bettis. 1986. The dominant logic: A new linkage between diversity and performance. *Strategic Management J.* **7**(6) 485–501.

PRNewswire. 2003. Energizer Holdings, Inc. acquires Schick-Wilkinson sword. (January 21).

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

Rumelt, R. P. 1974. *Strategy, Structure and Economic Performance*. Harvard University Press, Boston.

Rumelt, R. P. 1984. Towards a strategic theory of the firm. R. Lamb, ed. *Competitive Strategic Management*. Prentice Hall, Englewood Cliffs, NJ.

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 Sci.* **45**(8) 1109–1124.

Stigler, G. J. 1968. *The Organization of Industry*. Homewood, Irwin, IL.

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

Takayama, A. 1985. *Mathematical Economics*, 2nd ed. Cambridge University Press, New York.

Teece, D. J. 1980. Economics of scope and the scope of the enterprise. *J. Econom. Behav. Organ.* **1**(3) 223–247.

Teece, D. J. 1982. Toward an economic theory of the multiproduct firm. *J. Econom. Behav. Organ.* **3**(1) 39–63.

Teece, D. J., G. Pisano, A. Shuen. 1997. Dynamic capabilities and strategic management. *Strategic Management J.* **18**(7) 509–533.

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

Wernerfelt, B. 1984. A resource-based view of the firm. *Strategic Management J.* **5**(2) 171–180.

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

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

Winter, S. G. 1995. Four Rs of profitability: Rents, resources, routines and replication. C. A. Montgomery, ed. *Resource-Based and Evolutionary Theories of the Firm: Towards a Synthesis*. Kluwer Academic Publishers, Boston.

Winter, S. G. 2003. Understanding dynamic capabilities. *Strategic Management J.* **24**(6)(10) 991–995.

Winter, S. G., G. Szulanski. 2001. Replication as strategy. *Organ. Sci.* **12**(6) 730–743.

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

Yip, G. 1982. Diversification entry: Internal development versus acquisition. *Strategic Management J.* **3** 331–345.