Commensurability and collective impact in strategic management research:

When non-replicability is a feature, not a bug

Daniel A. Levinthal

The Wharton School of the University of Pennsylvania

dlev@wharton.upenn.edu

Lori Rosenkopf

The Wharton School of the University of Pennsylvania

rosenkopf@wharton.upenn.edu

November 14, 2020

We thank Drew Carton, Vit Henisz, Nicolaj Siggelkow, Batia Wiesenfeld, and attendees of the Wharton Management Department's virtual work-in-progress seminar for their comments. Rosenkopf acknowledges funding from Wharton's Mack Institute for Innovation Management.

Commensurability and collective impact in strategic management research:

When non-replicability is a feature, not a bug

Abstract

A persistent challenge in social science research is understanding whether and when empirical results generalize beyond a specific study's sample or context. In strategic management, "quasi-replication" examines whether results derived from particular industry, temporal, or geographic categories apply in adjacent research settings (Bettis et al 2016). We contend that the path to more robust and general theory must extend beyond quasi-replication by identifying the underlying factors driving both similarities and differences in results across research settings, which we call "basis variables." By transforming our usual categorical representations of research settings, basis variables promote commensurability, where seemingly distinct settings become comparable, enabling middle-range theorizing as theoretical contingencies are revealed. We close with suggestions to identify and elevate basis variables in individual and collective research efforts.

Scientists, social and natural, attempt to identify regularities in "nature" and to reveal the causal forces at work. The engine of progress in such efforts is some mixture of theory and observation. While observation inherently is situated in a particular set of circumstances of time and place, theory aspires to offer general and robust statements about the world. As a result, there is an inevitable gap between specific observation and broader theorizing, exemplified by Merton's (1947) call for theories of the middle range. For several decades, the heightened concern and energy around this mismatch in the social sciences has led several scholars to call for approaches that make our field's efforts more commensurable and impactful (Pfeffer 1994; Davis and Marquis 2005). There have also been calls for "quasi-replication" to assess how findings may or may not replicate in different research settings (Bettis et al 2016), reflecting concerns as to how firm some of our empirical building blocks might be, and even stronger reservations expressed about whether results do in fact replicate (Simmons and Simonsohn 2017).

We argue that the strategic management field will be well-served if we expand our emphasis beyond questions of replication or quasi-replication to focus our discussion on the *generalization* of research findings beyond individual studies. More specifically, when can findings observed via analysis of a specific sample in a specific context be assumed to hold in different settings? We suggest that such a collective research effort will hinge in important respects on achieving *commensurability* across disparate settings, by which we mean identifying common dimensions that enable comparability among separate categories. Achieving commensurability requires not only contrasting the findings themselves, but also grappling with the nature of the settings which generate these findings. This sensibility shifts the replication emphasis from the question of imperfect implementation that may affect results to the issue of

modifiers that might explain varied results across experimental or quasi-experimental settings. Thus, we call for more individual and collective activity to compare results over distinct settings and measures in order to identify important regularities across these efforts. We argue that a deeper understanding of the similarities and differences across our varied research settings is critical to creating more durable and applicable theories of the middle range, which in turn can yield greater impact on both academic and practical endeavors.

The challenge that we collectively face is identifying the dimensions of context that capture key causal factors that span particular settings. What may appear to be divergent settings as characterized by one representation of context, may be rather similar as characterized in a different representation space. Just as a strategy "landscape" may be viewed as "smooth" or "rugged" as a function of the strategist's cognitive representation (Gavetti and Levinthal, 2000), as a community of researchers we may face a patchwork of results that seem difficult to reconcile; but, potentially we may be able to identify a more coherent set of regularities by abstracting to higher-level construals of the constructs with which we situate populations of studies. Drawing an analogy to linear algebra, we posit that a judicious choice of "basis" variables by individual researchers and the broader field can help us generalize across the specific contexts we examine empirically, thereby progressing toward the art of mid-range theorizing as encouraged by Merton.[1] Ironically, the push for original theorizing in our empirical efforts arguably encourages treating results as having a universal quality --- and perhaps "overshooting" theories of mid-range to broad statements built on particular contexts. At the other extreme, if strategy research is not to fall prey to the critique of offering highly situated

[1] Basis vectors in linear algebra refers to a set of vectors that are independent and for which a linear combination of values along each of these vectors can characterize a point in the vector space. Bases are not unique, that is a different set of vectors might also satisfy this property, but each set would have the same dimensionality.

results, what Davis and Marquis (2005) pithily referred to as "journalism with regression equations," it is important to map the specific measures in individual contexts to a higher level of abstraction.

From replication to quasi-replication: Learning from empirical observation

While the concept of replication is a key principle of the scientific enterprise (Popper, 1959), the term "replication" has been applied to a variety of approaches and aims. Pure replication efforts literally reproduce prior studies. This may take the form of researchers recoding and reanalyzing archival data or "docking" a computational model onto a prior model. Experimental studies, however, vary across disciplines. In the physical sciences[2], environmental conditions and lab instruments can be controlled such that an experiment can be literally replicated. In the social sciences, however, true replication requires repeating the experiment with the same or nominally same population. The canonical example of such efforts is to take a classic experiment in psychology that was applied to an undergraduate population at University x at time t to a different undergraduate population at University y at time $t+\Delta$. Such an exercise is not literally a "replication", but rather a modest, in scope, examination of the generalizability and robustness of the initial findings. Do students at University y at time $t+\Delta$ respond to the stimulus of the experimental prompt in the same manner, up to some statistical measure of "sameness", as students in University x at time t?[3]

[2] We use the category "physical sciences" as we wish to exclude the life sciences, where in the biomedical context issues of generalizability across populations (gender, age, race, environmental conditions) have been proven important and challenging.

[3] One might naturally suggest a purer replication of the "test-retest" variety as consisting of re-examining the students at University x at $t+\varepsilon$, with ε as small as practically possible. However, the students at University x at $t+\varepsilon$ are now different than they were at time t. They have been subject to the experimental conditions and made certain responses to that stimuli, and as ε becomes less trivial in magnitude have been exposed to other stimuli in the world. In these respects, they are different from their prior selves when subject to the initial experiment. In that sense,

Replication efforts of this sort are at times motivated by some suspicion or concern about the validity of an existing, often highly cited, result. The implicit hypothesis in such research efforts is that if we were able to re-run history and treat the same subjects anew with the same experimental intervention, we might not get the same result (Aarts et al 2015). The (implicit) hypotheses for this non replication ranges from some suspicion about the treatment of the data, some intentional or unintentional extra-experimental priming by the experimenters, or the ethically neutral sense of a kind of sun spot anomaly related to idiosyncratic features of the initial experimental context.[4] Such efforts represent an important element of self-policing within the academy and are what tend to come to mind when we use the term "replication."

However, in the strategy field, so-called replication efforts vary in interesting and non-trivial ways with regard to the "subject pool". Our "subjects" are not convenient samples of college students, but typically consist of enterprises, often forming a population of firms in a specified industry at a particular time window. In this spirit, Bettis et al. (2016) distinguish between "quasi-replication" and pure replication. Since research contexts can vary by industry and by time, Bettis et al. argue that quasi-replication should only change one of these study dimensions at a time in order to allow for clearer comparisons across results.[5] As an example, Ghosh et al (2016) performed several quasi-replications by drawing from the Ahuja et al (2009) study of alliance formation in the chemical industry during the 1980s. Ghosh et al first quasi-

absent cloning or manipulation of memory ala the Matrix, pure replication is not possible in experimental social science.

[4] Per this issue of an "anomaly", a critical aspect of our empirical methods is what we take as a community to be a meaningful difference in results, with the current convention centered around a 5% chance probability that the result could stem from purely random fluctuations.

[5] Said differently, comparing the results of studies which may span different industries, time periods, methods and measures and attempting to infer what explains their differences and/or similarities would be a woefully under-identified activity.

replicated the focal study by analyzing 1990s chemical industry data with the same method. Next, they compared chemical industry results to semiconductor industry results during this same (1990s) timeframe. The key point here is that while pure replication would seek to confirm the same findings, excluding some statistical anomaly due to the "signal to noise" of the underlying pattern, quasi-replication efforts remove that expectation. Findings from adjacent contexts are considered with the proactive interest and intent that such efforts might yield distinct results and, in turn, raise the possibility of offering important theoretical insights as to moderating factors. Relatedly, Argyres et al (2020), in a special issue on the role of history in strategy research, argue that longitudinal archival data enable comparison of different temporal contexts to identify such moderating factors.

Another challenge with regard to the contrast of one set of results with another is the potential gap between construct and measures, a well-established stable of discussion within research methods (Bagozzi, Yi, and Phillips, 1991; Campbell and Fiske, 1959; Schwab, 1980).[6] As we examine research settings that vary in their context, even the nominally same measure may have quite different meaning. For instance, an "alliance" in the context of the software industry may mean something quite different than an alliance in biotechnology. In the former context, alliances are often marketing arrangements and may entail simply coordinating some of these downstream activities. In contrast, in biotech an alliance may represent a joint research effort regarding drug discovery or the clinical trials of some discovery. In the two settings, we

[6] We are distinguishing issues of measures from purely methodological examinations of robustness, which themselves can be important dimensions on which quasi-replications can be performed. Shaver's (1998) classic study of the mode of firms' entry ("greenfield" or acquisition) to new geographic markets demonstrates how using a more modern two-stage regression method to account for endogeneity dramatically alters our understanding of the implications of the choice of entry mode. This effort, what in hindsight can be seen as quasi-replication, established a new methodological standard for future work in our field. In a similar spirit, Villalonga (2004) shows that association between a financial discount and diversified firms was an artifact of not accounting for the endogeneity of the act of diversification itself.

have the same label, so at the semantic level we have "sameness", while the underlying phenomena may in fact diverge in important respects. While the issue of construct validity receives attention at times, particularly in the organizational behavior literature, the robustness and measurement of constructs across studies is less thoroughly considered in the strategy literature as measures such as software alliances are often highly situated in comparison to measures of individual person differences in a more psychologically-based study.

From results to generalization: Representing the research space and boundary conditions

While empirical findings demonstrate relationships between variables measured in specific settings, the extent to which these findings generalize to other settings is theoretical or speculative without additional empirical study. To portray this challenge of triangulating findings and interpreting the degree to which results generalize, we offer a more abstract representation of the research space in Figure 1. Empirical studies yield *results*, in the form of relationships between variables. These results are conceived as setting-specific, because the researcher has analyzed a specific *sample* within a particular *context*. The extensive range of phenomena examined in our field fosters a wide variety of sample-context combinations: examples might include a sample of chemical industry alliances formed during the 1990s; a sample of inventors patenting in the United States; a sample of CEOs from Fortune 500 firms; a sample of VC-financed startups during the tech bubble; or, a sample of multinational firms headquartered in rule of law countries. As such, samples are drawn at the level of individuals, firms, or partnerships, while contexts represent structural features that surround the actors, be they temporal, geographic, organizational, sectoral, or institutional.

Figure 1a depicts the range of potential empirical settings in the two-dimensional plane of sample x context. A single empirical study, represented by a point, generates result r as a function of its particular sample s and context c. While Figure 1a represents an individual study with the (s,c,r) triplet, Figure 1b adds a more general theoretical claim, depicting the expectation that results should generalize over broader regions of samples and contexts, as represented by the projection of a broad range of s x c settings (the parallelogram) onto a common result r. Thus, a particular empirical finding is but one test, for one sample-context combination, of a theory. The question of generalizability is whether the result holds in other samples and contexts, as suggested by the parallelogram. For simplicity, in Figure 1c we portray this question via local search along one dimension: a quasi-replication where context is held constant while variation in the sample s′ is examined, obtaining result r′.[7] If r′ is sufficiently close to r then the results may be viewed as generalizing across the two studies.[8]

Insert Figures 1a - 1c here

While quasi-replication compares two sample x context x result triplets, the broader problem, as illustrated in Figure 1b, is to identify the actual boundaries of a given set of relationships: what is the range of samples and contexts for which a given theory will hold? There may be empirical regularities -- truths -- across settings that are invariant over diverse

[7] Quasi-replication in context is analogous. As an example, the sample could be held constant, say a particular industry, and then the contrast would be made between two distinct time periods, perhaps over which there had been non-trivial technological or regulatory change.

[8] Of course, the tolerance around what is "sufficiently close" is a complex question in and of itself. In strategy research we may be satisfied to merely replicate the sign of a result. Other contexts may require obtaining a coefficient that is statistically equivalent.

samples (as in Figure 2a) or contexts, or even independent of both samples and contexts as in Figure 2b. Arguably, the most informative incremental empirical research effort is one that explores our beliefs regarding such boundaries --- under what settings does a given regularity start to break down?

Insert Figures 2a and 2b here

For the goal of constructing mid-range theory, identification of the boundaries at which the results do or do not generalize is a critical part of the research enterprise. If an empirical context is characterized by a particular organizational population at a particular point in time, it is not only of interest both to test whether a different sample or different context would generate a similar result, but also at least of equal interest is to discern when a distinct setting generates a different result. If we had a universal truth, such as the law of gravity, there would not be any bounds in the sample or context dimensions, as illustrated in Figure 2b. In this regard though, it is worth noting that even Newton's "universal" law of gravity was shown to in fact have bounds when Einstein's theory of general relativity demonstrated that the relationship Newton postulated breaks down for objects with extreme mass or at very close proximity. Einstein's theoretical breakthrough stemmed from a consideration of the boundary conditions and specifying his new theory required an understanding of these boundary settings.

The range of sample-context combinations may yield a more involved set of boundary conditions. Consider the stylized contrast posed by Figures 3a and 3b. In both figures, with three possible sample values and three possible context values, there are nine sample-context

combinations, five of which yield the result r, and four of which yield the result r'. However, whether the patterns of results are a useful catalyst to general theorizing depends on the order in which the rows and columns are presented. After all, how are we to linearly organize categories like industry, geography, occupation, and the like? The results as characterized by Figure 3a generate a rather confusing patchwork of seemingly highly contextualized results --- as one moves incrementally in either "s" or "c", the empirical results diverge sharply. However, Figure 3b takes these same nine combinations and permutes the rows and columns to yield a more coherent pattern of results. Again, it is important to note that the underlying empirical findings and associated settings have not changed, but rather how the different operationalizations of sample and context shift a seemingly rugged landscape of results to a smoother, patterned set of observations[9].

Insert Figures 3a and 3b here

Even with an ideal ordering along the axes, our theories may be more or less circumscribed, which in our conceptualization corresponds to a small domain of s and c values associated with a given r. While it is natural to think of a theory as being bounded by some range of s and c, boundaries and limits of theories may be more complex. In particular, we have depicted theories in Figures 1 and 2 as constituting essentially continuous functions that map sample and context to results. Intuitively, this implies that one can get arbitrarily similar results

[9] For a concrete example of this permutation, assume that the context axis is operationalized by industry. Merely ordering by categorical measures like SIC code is likely to engender more ruggedness than ordering them by underlying characteristics of the industry during the timeframe of the studies such as lifecycle stage, degree of appropriability, or concentration. Such re-orderings are likely to smooth the results landscape.

with a sufficiently proximate setting (sample and context).[10] However, the "patchy" pattern of 3a may not be a property of less than ideal characterization of the bases, but an inherent limitation of the theory. One might colloquially term such settings as a kind of "checkerboard" or "swiss cheese" landscape – where the limits not only define the maximal range of the space over which the theory applies, but those limits may be "interior" as well. That is, adjacent settings in the conceptual space may yield divergent results, while more distant "neighbors" yield similar empirical results. To offer a simple example of such a possibility from industrial organization, the issue of interdependence becomes negligible as the industry structure approaches a monopoly and again when the structure approaches the limit of perfect competition.

From categories to higher-level construals: Commensurability through basis variables

Traditional efforts at generalizing findings in our field tend to be more speculative than systematic, and typically focus on identifying similarities across settings rather than engaging differences. From a theory-building standpoint, we see generalizations offered in the concluding sections of papers when authors speculate about boundary conditions for their empirical findings, noting particular characteristics of industries (such as high-tech) or macro-economics (such as boom or bust cycles) where it may be reasonable to expect that similar results might, or might not, obtain. From an empirical standpoint, when data span multiple settings, we observe researchers including dummy variables for categories like industries or regions in their regression models, which separates setting-specific differences from the main effects rather than

[10] For a uniformly continuous function, there is for every given $\varepsilon > 0$ a $\delta > 0$ such that two values f(x) and f(y) have a maximal distance δ whenever x and y do not differ for more than ε. Thus, we can draw around each point (x, f(x)) of the graph a rectangle with height 2ε and width 2δ so that the graph lies completely inside the rectangle and not directly above or below.

exploring underlying factors that may have generated these differences (cf. Gulati and Gargiulo 1999).

Framing our interest as a question of generalization rather than merely quasi-replication clarifies that the question of interest is much broader than whether the same statistical relationships obtained in industry *x* at time *t* will emerge when examining industry *x* at time $t+\Delta$ or industry *y* at time *t*. Rather, it is more important to engage differences – to understand why a prior relationship no longer holds in a new setting. What it is about the two settings that may explain this divergence? As Johns (2006: 404) laments, "researchers seem almost desperate to ensure that reviewers and readers see their results as generalizable. To facilitate this, they describe research sites as blandly as possible --- dislocated from time, place, and space". Phenomenon-based, or more descriptive research, places context in the "foreground" of the work, while empirical efforts focused on testing theory tend to place context in the "background". Bridging this foreground and background, the research challenge is to speculate about the relevant explanatory contingencies.

While our representations of research space illustrate the challenges of identifying and smoothing the regions of samples and contexts over which results either generalize or diverge, these three-dimensional representations are stylized abstractions. Each dimension is depicted as a single ray in the figure, simplifying numerous attributes that define samples and contexts as well as results, and in that sense masks the challenge of comparing and contrasting results from different settings. For example, researchers in our field may sample at the level of individuals, organizations, or inter-organizational forms; furthermore, contexts may include temporal, geographic, organizational, sectoral and/or institutional factors. While sample and context are portrayed as orthogonal axes in our conceptual research space, the boundaries of these two axes

are far more fluid in reality, depending on the nature of the phenomena under study. As an example, samples of individual actors (like CEOs) may reside in organizational contexts (like Fortune 500 firms), while these same Fortune 500 firms might serve as the sample for a different study in other contexts (such as director interlocks in varied countries in the 1990s).

Our natural tendency to label samples and contexts using categorical variables – such as a particular industry, geography, or occupation – tends to situate us on more rugged research landscapes because of the specificity of these categories. In other words, a researcher might ask whether results from the semiconductor industry are generalizable to biotechnology, or to airlines, or to oil production. But the question of how to order these industry categories along a scalar axis is not straightforward, and the same issues arise for geographic or occupational categories as well. It is in this same spirit that Hernandez et al (2019) caution about the dangers of generalizing between individual and organization levels of analysis for network research.

Even the seemingly scalar context variable of time poses challenges. Whether study periods or measures are quantified in decades, years or days, it appears straightforward for a researcher to ask whether results generated from data spanning a particular timeframe (say the 1990s) are generalizable to earlier or later timeframes (decades). Furthermore, while it is common to use time as a possible boundary condition in strategy research, time in and of itself is not typically the critical concern. The year 1980 is not different from 2000 merely because twenty years of time has passed. Rather, the possible contrast between the empirical properties we might observe in the two periods stems from radical changes in technology, the information technology revolution in computing both in the form of computation devices and their linkage via the Internet, globalization, deregulation, and/or economic conditions. As Johns (2006: 389) notes, time is a surrogate for the environmental stimuli at the time the research is conducted.

Thus, even "time" is analogous to a categorical variable (1980s versus 1990s versus 2000s), proxying multiple changes in the context across periods. Importantly, this implies that if we had direct measures of the relevant contextual attributes, time would no longer be a useful dimension.

For example, when Ghosh et al (2016) found that prior alliances between firms in the chemicals industry positively predicted their subsequent alliances in the 1980s but negatively predicted them in the 1990s, they speculated that the underlying factors reversing this taken-for-granted result in the 1990s context were the maturity of the industry and its concomitant consolidation of firms. Moreover, abstracting from specific settings like industry-time period combinations to higher-level construals like industry maturity or concentration provides commensurability across empirical settings that may otherwise seem incommensurable: chemicals in the 1980s and semiconductors in the 2010s differ in both temporal and industry space, but may be more comparable in terms of lifecycle stages.

In referring to these higher-level construals as *basis variables*, we borrow this term from linear algebra as referring to the dimensions that characterize a given space. Basis variables offer an alternative from broad categorical variables, such as time period and industry. Furthermore, per our Figure 3a and 3b contrast, appropriately chosen basis variables can generate relatively smooth research landscapes that would otherwise have appeared more rugged when operationalized over specific, categorical axes like industries and decades. Thus, higher-level insights can stem from the abstraction of industry categories to a wider set of commensurable constructs such as stage of industry lifecycle, pace of technological change, minimum efficient scale, ecosystem complexity, and the like. These constructs transcend the specific industry categories and therefore provide more insight about specific industries not yet studied. Likewise, comparing findings from similar studies across different occupations becomes more deeply

meaningful when we abstract from occupational categories to constructs like job complexity, clarity of output, interdependence with other occupations, and the like. Or, considering findings across communities of academics or practitioners, the nature of the knowledge – dispersion of expertise, pace of change or growth, interdependence among components – may provide this more abstract guidance.

How do we find appropriate candidates for basis variables? Theories provide guidance when we examine the underlying engines of their predictions. For example, evolutionary and ecological perspectives rely on history and population demography; transaction cost and network perspectives rely on dependence, concentration and constraint. RBV perspectives highlight appropriability while institutional perspectives highlight legitimacy. Abstracting from these labels, we offer a candidate set of three high-order constructs – uncertainty, interdependence, and demography – that can guide researchers in thinking about specific basis variables most relevant for their school of thought. Uncertainty refers to the extent to which attributes in a given time period are predictive of attributes in the next period; interdependence refers to relational characteristics among actors in the research setting; and demography refers to the distributions of attributes across actors.

Table 1 suggests basis variables appropriate for the analysis of firm-level behavior and performance.[11] The entries in each cell represent underlying characteristics of categorical research settings (denoted by rows) that convey the overarching constructs of uncertainty, interdependence, and demography (denoted by columns). Specifically, industry categories may be more usefully represented by measures conveying uncertainty, such as rate of technological change, levels of standardization or modularity, appropriability or legitimacy. Continuing with

[11] Of course, analogous tables may be built for individual or inter-organizational units of analysis.

measures conveying the demography of an industry setting, measures of density (that is, population size), minimum efficient scale, and market concentration can capture implications for competitive intensity, as can first and second moments of measures connoting resources and capabilities derived from patent stock, cumulative expenditure on R&D, or size-based measures of various attributes such as sales force or physical plant and equipment. Finally, considering how interdependence may manifest across industry settings leads us to propose constructs like value chain position (as conveyed by input-output tables) as well as the network structure among actors in the industry.

Basis variables that connote underlying characteristics of regions can also be useful. Once again, using the themes of uncertainty, demography and interdependence, we can suggest several candidates. With respect to uncertainty, key institutional characteristics such as regulatory and political regimes are likely to shape results, as well as the level of economic development. Interdependence may be captured by trade balances and measures of the maturity and munificence of the ecosystem. For demography, considerations of technological specialization and cultural factors are also relevant. Here, we should acknowledge that multinational scholars have historically grappled with the translation of country labels to basis variables. creating indices that aggregate numerous measures into broader constructs like culture (Hofstede, 2002) and political constraint (Henisz, 2000). Likewise, work on comparative management (Guillen, 1994; McDermott, 2002) has contrasted settings, usually on the basis of economic or political system, making inferences on the basis of these broad system-level differences.

As we have noted previously, these characteristics of industries and regions are time-varying; using them as basis variables reduces the explanatory power of time period as a

categorical variable. But at the same time, it is useful to note several time-varying characteristics that are distinct from the research settings of particular industries or regions. Here, macro-economic conditions can provide additional insight regarding uncertainty, just as the development of enabling technologies may impact interdependence or managerial trends in organizational forms and practices may shape demographic considerations.

Insert Table 1 here

From generalization to impact: Implications for strategic management research

Our call for emphasizing basis variables to build commensurability across research settings serves as a counterbalance to the tendency for work to operate at one of two "poles" of either treating specific contexts as if they are representative of a broad array of settings without direct testing or consideration of that implied claim or viewing a specific instance as a "unicorn" and unique to that context. Operating in the interior of these two poles and building a more cumulative body of work may require aligning the incentives and motivations at the individual researcher level with collective aims for the field.

Strategy research in historical context. When viewing the evolution of research in strategy, we can see the pendulum swinging between these two poles. Recall that the founding approaches of strategic management, rooted in the field of IO economics, explored how variation in industry structure (context) shaped firm conduct and performance (Bain, 1968; Scherer, 1970). Here, large panel datasets derived from SEC filings and aggregated sources like COMPUSTAT served as community assets, generating natural connections between studies. Over time, the

pendulum swung away from efforts of this type due to two critiques. First, industry boundaries were "taken for granted" in the categories defined by the regulatory structure; however, this was problematic as such classifications are often mismatched against what would be appropriately regarded as a set of firms providing close substitutes. Second, the findings derived from cross-sectional or panel data sets generally neglected the critical incipient stage of an industry's development when many of the participants are small, privately-held enterprises.

As a result, the pendulum swung toward rich, industry-specific, "hand-collected" data, emphasizing new lines of work within the literatures on evolutionary economics, population ecology, and technology management (Carroll and Hannan, 1989; Klepper and Grady, 1990; Klepper and Simons, 2005; Tushman and Anderson, 1986; Agarwal and Gort, 1996). These scholars built impressive detailed industry histories over a number of distinct industries and were able to identify regularities of industry dynamics (such as density dependence, shakeouts, and responses to technological change) that spanned these different settings. Much of this effort was directed at identifying stable patterns that occurred across a range of industries. In some cases (Klepper and Grady, 1990; Agarwal and Gort, 1996), the work itself spanned multiple industries, while in population ecology the establishment of robust patterns occurred over a set of industry specific studies (Carroll and Hannan, 1989). However, these lines of inquiry were generally not focused on identifying the boundary conditions of these "regularities". An interesting exception in this regard is the work of Tushman and Anderson (1986) that examined technological change across a set of industries (cement, glass, and minicomputers), demonstrating that the nature of this technological change (competence-enhancing or competence destroying) provided key context for its effects.

Indeed, while these proprietary data yielded rich and novel insights, this approach arguably exacerbated the challenge of generalizing across contexts because these data sets were generally researcher-specific assets rather than "community property" of the broader academic community. In an effort to overcome the challenges and limitations of independent development of proprietary datasets, Helfat and collaborators (the FIVE project) have tried to orchestrate a community-level resource of shared industry data sets to facilitate opportunities for comparative analyses above and beyond the strengths of the richness and completeness of these carefully constructed industry datasets. Such efforts can provide an important public good, though their creation have the natural challenges generally associated with knowledge management systems around the value of private "knowledge" and the individual costs of codifying this private knowledge in a way that is accessible to others (Hansen and Haas, 2001).

Data sources for testing basis variables. Fortunately, the growing availability of "big data" may offer new opportunities for engaging in studies across populations which would have traditionally been examined as separate research settings. For instance, Rosenkopf and Schilling (2007) use SDC alliance data to explore network structure in over 30 industries, and Schilling (2015) examines alliance formation across the universe of alliance data over 20 years. With more extensive archival databases like these, there are clear opportunities for researchers to test their industry-specific findings across a wide range of industries, coding potential basis variables like industry lifecycle or type of technology and examining their effects across time and space. Such approaches would strengthen our theory by providing rich context variables, as opposed to a reliance on industry dummies, because such measures are comparable across a range of settings – the very essence of middle-range theory.

While past efforts developing panel datasets were typically built on the back of the artifacts of the regulatory system around financial accounting systems (e.g., COMPUSTAT), government recording keeping (e.g., census of manufacturing) and specific commercial interests of analysts (e.g., SDC platinum or Pitchbook), the modern internet era offers new possibilities from crowd-sourced data. Glassdoor has been used to assess organizational culture (Marchetti, 2020). LinkedIn has been used in a number of papers to capture individual skills (Tambe and Hitt, 2013). These web-based aggregators are providing a transparency at the level of organization populations that we previously had lacked.

While such efforts provide exciting opportunities to consider a wide variety of settings using reasonably comparable measures, they also pose their own challenges. For example, while LinkedIn provides rich micro-level data on individuals' careers, creating a correspondence from those very specific sets of experiences (roles and industry context) to more general measures still requires agreement across researchers in order to achieve commensurability. Even if we have what is nominally the same measure, for instance someone in the role of CFO, what that role means behaviorally across settings may differ considerably. For instance, a CFO may have a narrow focus on financial management and reporting or their role may include broader considerations such as corporate development and mergers & acquisitions. Thus, a given measure can take on different meanings in different settings. Perhaps even more challenging is to go from these micro level data to a construct of "industry" and, more to the point, the distinct underlying "macro" context in which these actors are operating.

Techniques for inducing basis variables. As we have argued, identifying basis variables that can make seemingly different contexts more commensurable is the key to research progress. Work that examines the role of potential basis variables across a variety of theoretical ideas and

conceptual operationalizations and provides higher-level construals with which to represent research settings could help reconcile disparate extant findings resulting from more specific operationalizations. Said differently, when faced with conflicting or negative findings, one mode of response is to specify an alternative theoretical mechanism. A more incremental approach might be to accept as valid the prior findings and examine what factors appear to be creating the boundary between the prior findings and the current one.

Fortunately, in addition to identifying basis variables from extant theory and empirics, new work rooted in machine learning holds the promise of identifying key factors computationally. Topic modeling can delineate categories with greater similarity; yet, the underlying meaning of these machine-generated categories is often not well-understood. Human intelligence is needed to interpret these outcomes, just as it is needed to interpret across simpler quasi-replications.

Qualitative research can also play an important role by providing a powerful "search engine" to identify basis variables via comparative case exercises. Even single case qualitative studies can provide additional stimuli for future research by illustrating anomalous situations where expected results for basis variables do not hold. Qualitative Case Analysis (Ragan, 1989) is an interesting middle-ground between the classic approach of comparative analysis and the shift to the direct measure of underlying constructs across settings. Key to the power of QCA is that attributes, due to properties of complementarity and substitution, may be co-occurring in the former case or exclusionary in the latter situation. QCA is generally applied when the researcher has a moderate number of cases, too many for a direct comparison of cases but perhaps too few for some statistical analysis of underlying attributes --- though as we note this latter challenge may be as much our difficulty in specifying and measuring the critical underlying dimensions as

a limitation of the sample size. QCA can serve as a useful inductive, exploratory tool in our efforts to identify more or less useful bases by which to dimensionalize our conceptual space.

It is also worth noting that our emphasis on generalization through basis variables bears resemblance to meta-analysis. Traditional meta-analytic efforts in the management domain are more common in the micro-organizational behavior literature than on the macro side (Ostroff and Harrison, 1999), and we suspect this is largely due to the greater commensurability of studies that utilize well-established instruments for psychological constructs. Such meta-analyses are not common on the "macro" side given not just a limited number of data sets, but also the challenge of commensurability of measures across industry settings.

Recent work on random coefficient models (Alcacer et al., 2018) points to a useful way to systematically identify settings where our usual regression approaches might mask a richer and more contextualized pattern of behavior and results within a given sample. As Alcacer et al. note, a standard regression approach identifies average effects over a sample population. Therefore, unmodeled contingent factors or characteristics of the individual firms might yield mixed or null findings in spite of the possible presence of systematic properties. In contrast, the random coefficient approach captures the variance in the parameter estimate as well as the average value. Thus, an estimate with an average effect that is estimated to be close to zero but with a large variance suggests that, within the sample, firms' performance outcome is determined in a very different manner. As they note (Alcacer et al., 2018: 534), "researchers can then use distributional estimates as a jumping-off point to explore the reasons behind firm heterogeneity."

Similarly, it is also important to remain mindful of how we define and limit samples in a given research setting. Issues of the non-representativeness of samples are gaining attention, whether concerns of race and gender emerging from machine learning techniques for facial

recognition and other applications, or the U.S.-centric nature of a vast set of strategy research. Researchers are now very conscious of sample selection effects in the form of the possible endogeneity of what they might be tempted to treat as an exogenous variable, such as organizational structure or strategy. However, we as a community have been relatively acquiescent about the convenience, or saliency-based, sampling in which we as researchers engage. Industries in which patenting is a meaningful marker of knowledge are substantially over-sampled as we as researchers very much value this measure of an otherwise difficult to ascertain attribute. However, we know that industries for which patents are a critical device to appropriate knowledge are relatively rare and presumably are non-representative on other dimensions as well; yet, for all our concern for endogeneity of behavior within a sample, we tend to give ourselves as researchers a collective pass.

Institutional (field-level) efforts. Since efforts to standardize measures underlying our main conceptual categories can inform dataset development and subsequent comparisons across studies, we implore institutional gatekeepers to elevate efforts and construct incentives to build consensus around key basis variables. Review papers and special issues focusing on antecedents and consequences of candidate basis variables can inspire more systematic vocabulary and measures across seemingly disparate research settings. Furthermore, editors can encourage authors to connect their data and theorizing to extant basis variables. In other words, requiring attention to scope conditions in journal submissions linking the paper's research setting to extant basis variables aids the quest for commensurability. Likewise, collaborative projects extending quasi-replications across varied research settings can also identify boundary conditions; this approach may be all the more critical where primary data is proprietary and built through intensive organizational study (cf. GLOBE project (Dorfman et al., 2012)).

Conclusion

Theorizing in strategic management is far more bounded than that of Newtonian and Einsteinian physics. Social entities, such as firms and industries, are likely to be situated in more complex ways than atoms and planets. This is a limit of theorizing in the strategy domain, but at the same time a source of its richness: the greater the frequency of boundaries, the greater the opportunity to refine and embellish mid-range theorizing. As our field has traditionally prized novelty of findings over replicability of findings, it is unsurprising that individual researchers pursued novelty through idiosyncratic constructs and measures at the expense of testing the scope of extant theories via quasi-replications.

Mapping the research space effectively requires moving beyond single or pairwise comparisons to broader populations of results. We must look for boundaries and breaking points across research space, not to suggest that prior scholars misinterpreted their own data, but to illustrate that the relationships suggested by these prior findings are more circumscribed than perhaps previously thought or understood. Ideally, such efforts will serve as a catalyst to new theorizing. A useful set of quasi-replications, creating seeds for mapping research spaces, have been set forth in the Strategic Management Journal's special issue on replication in strategic management (2016). SMJ's ongoing encouragement for the publication of quasi-replications (Ethiraj et al 2016) is a valuable institutional spur that should serve as a model for other top journals in the management orbit. Advances in data science also enable new approaches with wider scope (Puranam, 2019). However, our path forward will be facilitated not just by technological advances in data science, both the proliferation of data sets and the algorithms to

explore them, but by our understanding of the fundamental challenge of extracting some possible relatively general truths out of the restricted sample of our collective experience.

References

Aarts, Alexander A et al 2015. "Estimating the reproducibility of psychological science." *Science* 349(6251).

Agrawal, R. and M. Gort (1996). "The evolution of markets and entry, exit and survival of firms". *Review of Economics and Statistics,* 78(3): 489-498.

Ahuja G, Polidoro F, Mitchell W. 2009. Structural homophily or social asymmetry? The formation of alliances by poorly embedded firms. *Strategic management journal* **30**(9): 941-958.

Juan Alcácer, Wilbur Chung, Ashton Hawk, Gonçalo Pacheco-de-Almeida (2018). Applying Random Coefficient Models to Strategy Research: Identifying and Exploring Firm Heterogeneous Effects. Strategy Science 3(3):533-553.

Argyres N, De Massis A, Foss NJ, Frattini F, Jones G, Silverman BS. 2020. "History-informed strategy research: The promise of history and historical research methods in advancing strategy scholarship." *Strategic Management Journal* 41(3): 343-368.

Richard P. Bagozzi, Youjae Yi and Lynn W. Phillips. Assessing Construct Validity in Organizational Research Source: *Administrative Science Quarterly*, Vol. 36, No. 3 (Sep., 1991), pp. 421-458.

Bain, J.S. *Industrial organization* (2nd ed.). New York: Wiley, 1968.

Bettis, Richard A., Helfat, Constance E. and J. Myles Shaver. 2016 "The necessity, logic and forms of replication." *Strategic Management Journal*, 37(11): 2193-2203.

Campbell, Donald T., and Donald W. Fiske 1959 "Convergent and discriminant validation by the multitrait-multimethod matrix." *Psychological Bulletin*, 56: 81-105.

Carroll, G.R., Hannan, M.T., (1989). "Density delay in the evolution of organizational populations: a model and five empirical tests". *Administrative Science Quarterly* 34 (3), 411–430.

Choi, J. and D. Levinthal (2019). "Wisdom in the wild: Generalization and adaptive dynamics".

Cook, T. and D. Campbell (1979). *Quasi-Experimentation: Data Analysis in Field Settings*. Chicago, IL: Rand McNally College Publishing.

Davis GF, Marquis C. 2005. Prospects for organization theory in the early twenty-first century: Institutional fields and mechanisms. *Organization Science* **16**(4): 332-343.

Dorfman, P., M. Javidan, Pl Hanges, A. Dastmalchian, and R. House (2012). Globe: A twenty year journey into the intriguing world of culture and leadership. *Journal of World Business*, 47: 504-518.

Ghosh A, Ranganathan R, Rosenkopf L. 2016. The impact of context and model choice on the determinants of strategic alliance formation: Evidence from a staged replication study. *Strategic Management Journal* **37**(11): 2204-2221.

Guillen, M. (1994). *Models of Work: Work, Authority, and Organization in Comparative Management*. University of Chicago Press, Chicago IL.

Gulati R, Gargiulo M. (1999). Where do interorganizational networks come from? *American Journal of Sociology,* 104(5): 1439-1493.

Hansen, M.T., M.R. Haas. 2001. Competing for attention in knowledge markets: Electronic document dissemination in a management consulting company. *Administrative Science Quarterly* **46**(1) 1-28.

Henisz, W. (2000). "The institutional environment for economic growth". *Economics and Politics*, 12: 1-31.

Hernandez E, Kleinbaum A, Shipilov A. 2019. A Network Is A Network? (Non)-Generalizability across Levels of Analysis in Network Research.

Hofstede, G. (2002). Culture's Consequences: Comparing Values, Beliefs, Institutions and Organizations across Nations. Sage Publications: Thousand Oaks, CA.

Johns, G. (2006). "The essential impact of context on organizational behavior". *Academy of Management Review,* 31(2): 386-408.

Klepper, S. and K. Simmons (2005). "Industry shake-out and technological change". *International Journal of Industrial Organization*, 23: 23-43.

Klepper, S. and E. Graddy (1990). The evolution of new industries and the determinants of market structure. *RAND Journal of Economics:* 21 (1), 27– 44.

Lavie D, Rosenkopf L. 2006. Balancing Exploration and Exploitation in Alliance Formation. *Academy of Management Journal* **49**(6): 797-818.

Marchetti, A. (2020). "Firms of a feather flock together: The role of acquirer-target culture compatibility in technology acquisitions". INSEAD Working Paper.

McDermott, G. (2002). *Embedded Politics: Industrial Networks and Institutional Change in Post-Communism*. University of Michigan Press, Ann Arbor, MI.

Ostroff, S. and D. Harrison (1999). "Meta-analysis: Level of analysis, and best estimates of population correlations: Cautions for interpreting meta-analytic results in organizational behavior". *Journal of Applied Psychology*, 84(2): 260-270.

Pfeffer, Jeffrey. 1993. "Barriers to the advance of organizational science: Paradigm development as a dependent variable." *Academy of Management Review.* 18(4).

Popper, K. (1959). *The Logic of Scientific Discovery*. Routledge, NY.

Puranam, P (2019). "Is the theorist an endangered species". Journal of Marketing Behavior, 4(1): 43-81.

Rosenkopf L, Schilling M. 2007. Comparing alliance network structure across industries: observations and explanations. *Strategic Entrepreneurship Journal* **1**(3-4): 191-209.

Scherer, F.M. (1970). *Industrial Market Structure and Economic Performance*. Chicago: Rand McNally.

Schwab, Donald P. 1980 "Construct validity in organizational behavior." In L. L. Cummings and B. Staw (eds.), Research in Organization Behavior, 2: 3-43. Greenwich, CT: JAI Press.

Shaver, J. Myles. 1998."Accounting for endogeneity when assessing strategy performance: Does entry mode choice affect FDI survival?" Management Science 44 (4): 571-585.

Schilling, Melissa. 2015. "Technology shocks, technological collaboration, and innovation outcomes." Organization Science 26(3): 668-686.

Simmons, Joseph P. and Uri Simonson. 2017. "Power posing: P-curving the evidence." Psychological Science 28(5): 687-693.

Tambe P, Hitt LM (2013) Job hopping, information technology spillovers, and productivity growth. *Management Sci.* 60(2): 338–355.

Tushman, Michael L and Philip Anderson. 1986. "Technological discontinuities and organizational environments," Administrative Science Quarterly 31(3): 439-465.

Villalonga, Belén (2004), "Does diversification cause the 'diversification discount'?," Financial Management 33 (2), 5–27.

Figure 1a: Conceptual view of research landscape

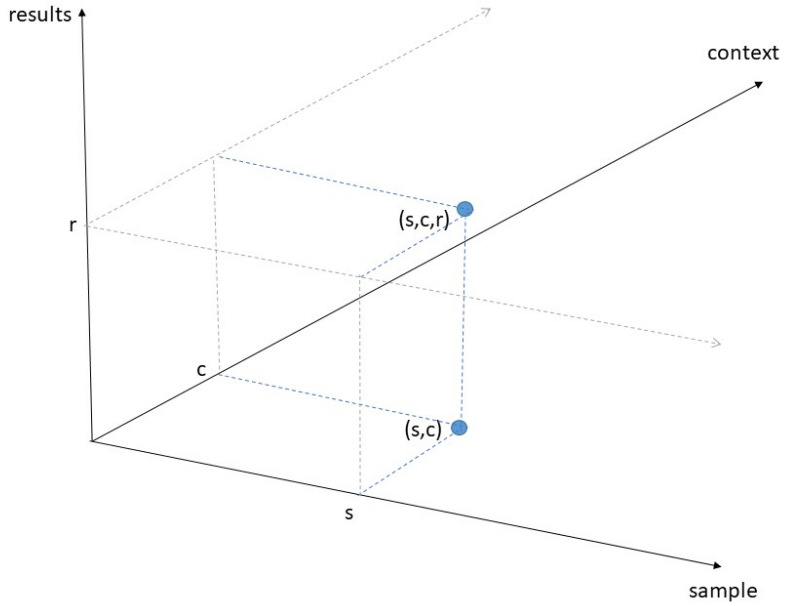

Figure 1b: Theoretical claim spanning a range of samples and contexts

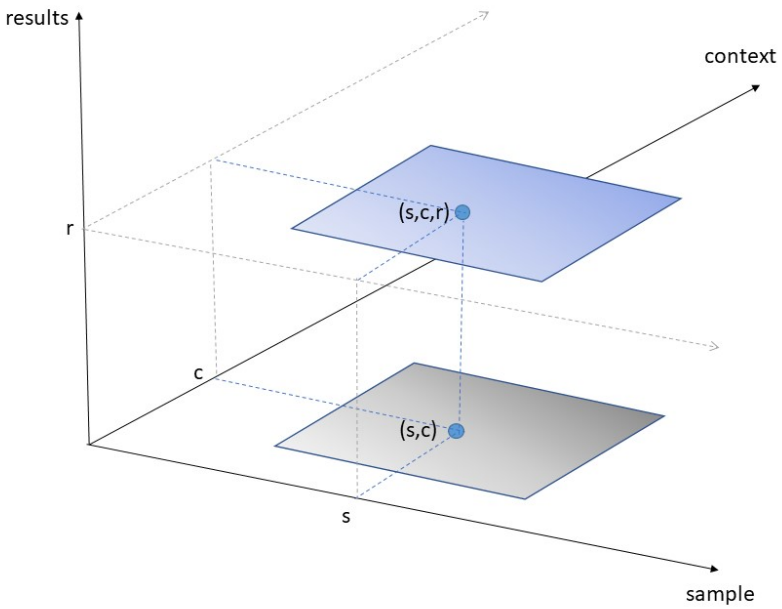

Figure 1c: Quasi-replication across different samples (same context)

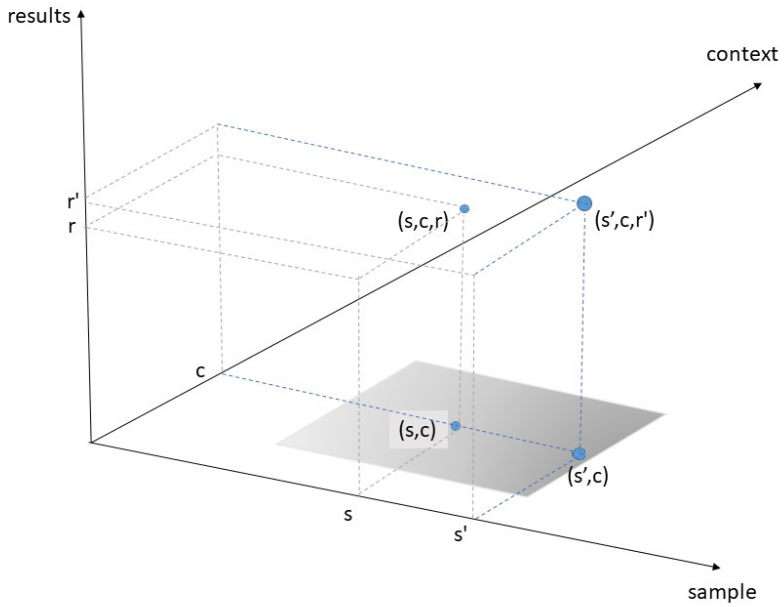

Figure 2a: Sample-invariant truth

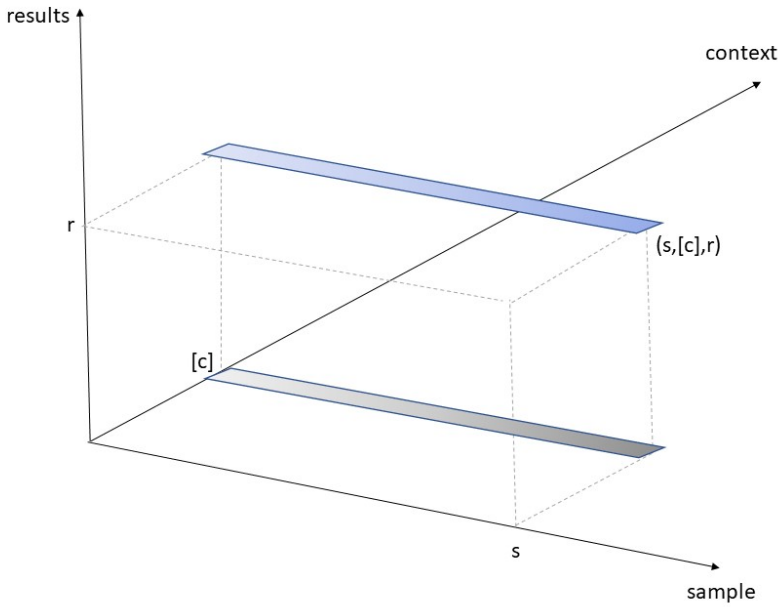

Figure 2b: Empirical regularity

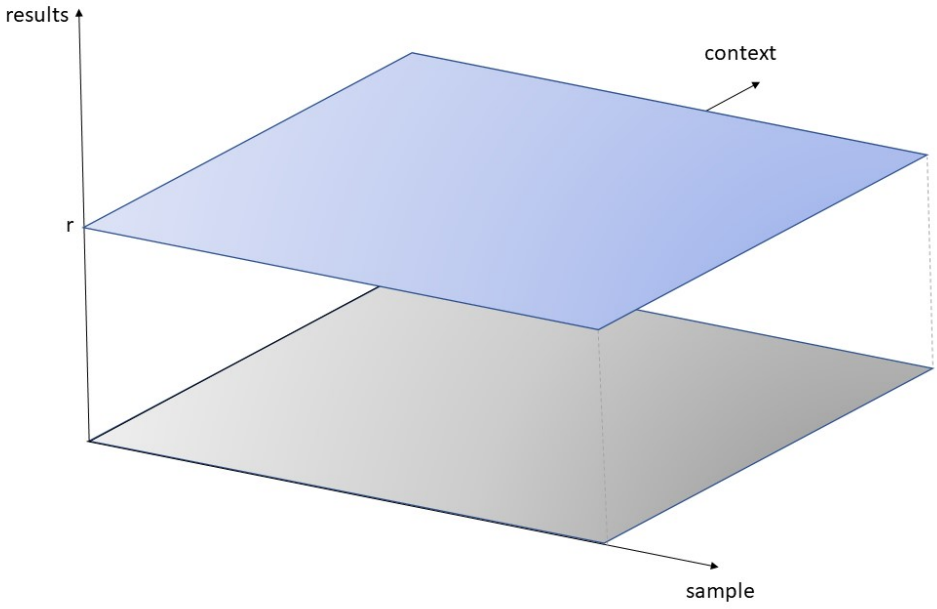

Figure 3a: “Rugged” landscape of empirical results

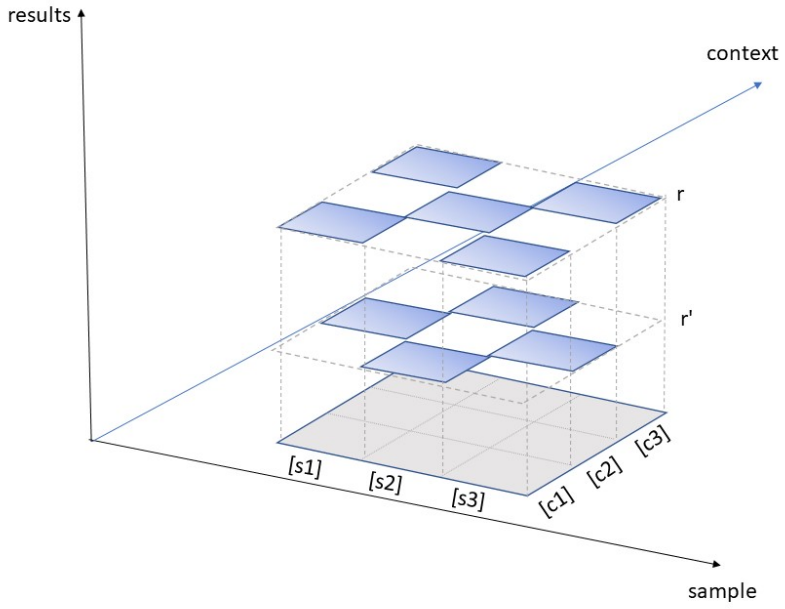

Figure 3b: Transformed basis and “smooth” pattern of results

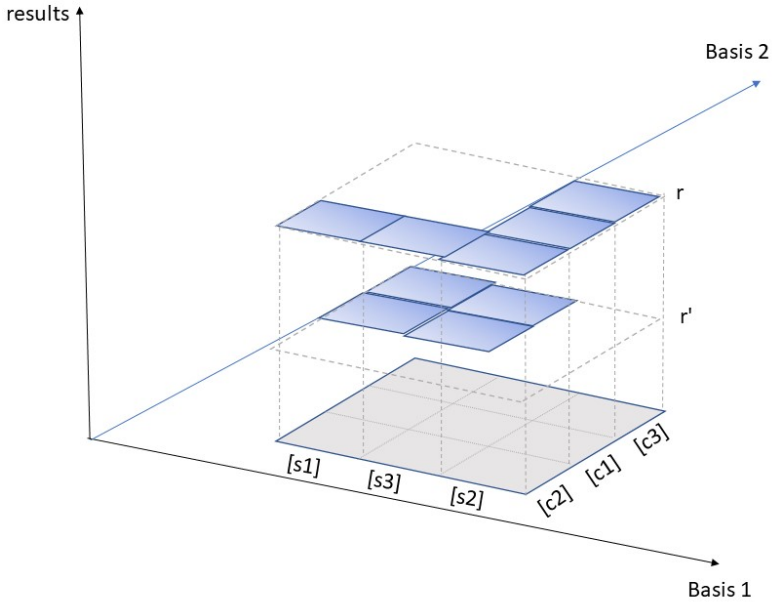

Table 1: Candidate basis variables at the firm level of analysis

		Overarching themes across research settings		
		Uncertainty	**Demography**	**Interdependence**
Categorical research settings	**Industry**	Rate of technological change Standardization/modularity Appropriability Legitimacy	Market concentration Density Minimum efficient scale Resources / capabilities	Value chain position Network structure
	Region	Regulation Political system State of economic development	Tech specialization Culture	Ecosystem development Trade balance
	Time period	Macroeconomic conditions	Managerial practice	Enabling technologies