Organization Science (in press)
Brice Dattée and James Barlow
https://doi.org/10.1287/orsc.2017.1113

Multilevel organizational adaptation:
Scale invariance in the Scottish healthcare system

Organizational adaptation results from coupled processes nested across multiple levels (Klein and Kozlowski 2000; Klein et al. 1999). There has been a strong call for a quasi-natural organization science that accounts for the effects of coupled processes across different levels (Lewin and Volberda 1999; McKelvey 1997). Yet, most work on organizational adaption has been conducted implicitly at a single or dual level of analysis. To remedy this, detailed empirical cases with more than two levels of analysis are required to address the interactions between organizational context and the intra-organizational dynamics (Greenwood and Hinings 1996; Tracey et al. 2011) .

We explore how the classical problem of limited cognitive representation influences the way agents coordinate their search efforts across organizational boundaries and across multiple levels. We combine two established perspectives and apply them recursively across multiple levels: the cognitive and experiential search efforts of unitary agents in complex settings (Gavetti and Levinthal 2000; Levinthal 1997; Levinthal and Warglien 1999; Rivkin and Siggelkow 2007), and the phenomenology of coordinated search situated in hierarchical organizations (Gavetti 2005; Knudsen and Srikanth 2014; Siggelkow and Rivkin 2005). Specifically, we explain how complex cooperation in hierarchical systems can emerge and subsequently unravel. We show how cognitive limitations can hide opportunities for collaboration and mutual improvement. We demonstrate the multilevel dynamics induced when agents at various levels broaden their cognitive representations to include higher-order epistatic interactions – i.e. interdependencies between components which affect system performance – and synchronize their adaptive search across organizational boundaries. Our demonstration is based on a longitudinal case of a large change programme in a national healthcare system, where there was a powerful rhetoric on the need for "whole system change". This aimed to make agents in all its constituent parts aware that they are interdependent and to emphasize the need for collaboration to uncover opportunities for mutual performance improvements:

> *"This new approach is about getting the NHS [National Health Service] in Scotland to work as a single, whole system. We need all of the partners in the system to realize that they are interdependent. Action in one part of the system has an impact elsewhere. And we need the partners to understand that we all need to change."* NHS Scotland, Policy document, 2005

Agency theory models take the structure of interdependencies as a given then focuses on incentive alignment among rational actors to guide local action towards an identified global peak. The literature on organizational adaptation, however, assumes that individual behavior is driven by adaptive search processes, so that cooperative behavior among self-interested individuals may be induced by manipulating the set of interdependencies (Levinthal and Warglien 1999): managers at all levels must trigger cascades of cooperative behavior and influence adaptive actions across levels and boundaries (Gavetti and Levinthal 2000; Tsoukas and Chia 2002; Weick 1995).

Coordination of joint-search in organizations

Hierarchical systems can be decomposed into interactions within and between subsystems (Simon 1962). Within a subsystem, the local adaptive search process of agents depends on the topography of the payoff surface in which they are embedded – the peaks and troughs they must negotiate to achieve efficient management (Kauffman and Levin 1987; Levinthal 1997) – and on their cognitive representations of action-outcome linkages in their particular problem space (Gavetti 2005; Gavetti and Levinthal 2000). Prior work, however, has three related limitations. First, models of organizational search are essentially *non*-organizational, assuming a single agent and ignoring the epistemic interdependencies involved in multi-agent search and coordination (Knudsen and Levinthal 2007). Apart from (Puranam and Swamy 2016), the concept of coupled learning during joint search is entirely absent, even if models in game theory do take into account epistemic interdependence (Knudsen and Srikanth 2014). Second, the literature considers that agents know that they are interdependent and on which dimensions, then engage in joint search. Even though Puranam and Swamy (2016) acknowledge that the precise nature of interdependencies may be initially unknown to agents, the assumption that agents are actually aware they are interdependent remains. Third, attention has been on the coupling dynamics across a given boundary, which implies that these analyses are always conducted at a single level of a system. Siggelkow and Rivkin (2005) attempt to handle multilevel dynamics, whereby department heads may be required to send change proposals to a higher layer of management for arbitration. Yet, their model assumes that firm-wide incentives, which they show reduce the likelihood of mutually destructive efforts on the part of department heads, are reciprocal.

This assumption of reciprocity may have important implications for organizational adaptation, especially if agents belong to more remote, but nonetheless interdependent, parts of a large organization. Management can present a sense of what constitutes higher performance, and enable agents to jointly search around the most attractive options (Levinthal and Warglien 1999). However, agents develop their own cognitive representations that simplify even their immediate organizational structure (Gavetti and Levinthal 2000), possibly hiding opportunities for collaboration to address the inevitable conflicts inherent in organizations. Collaboration implies a risk of loss, hence there must be benefits to achieve coordinated collective action. Broadening cognitive representations to discover higher-order interdependencies may be challenging for some agents. The current coupling approach assumes either that the idiosyncratic cognitive representations of agents overlap with those dimensions enabling coordinated search across boundaries, or that cognitive shifts are instantaneous.

To address these limitations, we draw on complexity theory applied to organization science (Allen et al. 2011; Anderson 1999; Anderson et al. 1999; Eisenhardt and Piezunka 2011; Maguire et al. 2006; Thietart and Forgues 2011). Complexity theory guides us to focus on multilevel interactions, dissipation across boundaries, and time and scale. We extend the theory of organizational adaptation by establishing scale-invariant and scale-dependent results from a detailed empirical case study.

Managing in a complex system

Although a common criticism is that it is overly metaphorical and hard to apply practically to management, the popularity of complexity theory in organization science has grown, both in research and in management and policy rhetoric (McKelvey 1999; Moldoveanu and Bauer 2004; Zhou 2013). Literature stresses the importance of recognizing non-linear dynamics within a complex system and the recursive nature of organizational activities which are similar, repeated, and mutually defined across levels (Andriani and McKelvey 2009). The role of a manager under the complexity paradigm is to manage twin tensions: a horizontal tension for coordination across boundaries and a vertical tension for self-organization. First, managers need to facilitate local experimentation within the boundaries of the system over which they have direct control, while also influencing and coordinating with the outside networks across boundaries. This creates a horizontal tension, “landscape coupling” (Levinthal and Warglien 1999). When exerting influence across an open system’s boundaries, complexity theory

alerts us to the inevitability of dissipation, that is energy or effort converted to no benefit; hence the need to sustain efforts to compensate until the effect of changes accumulate over a certain threshold (Anderson 1999). However, sources of dissipation are not comprehensively addressed in the organizational adaptation literature.

Second, managing in a complex system requires balancing hierarchically imposed top-down rules that guide local actions, while at the same time creating the conditions for enabling bottom-up emergence. This adaptive tension to maintain a semi-structure creates the conditions for self-organization. Radical shifts may result from the accumulation of frequent and small changes in semi-structures (Brown and Eisenhardt 1997), and from agents revisiting their mental models of their interdependencies (Boisot and Child 1999; Gavetti and Levinthal 2000; Tsoukas and Chia 2002). By devolving authority and empowering agents, self-organization requires us to accept "equifinality", the principle that a given end-state may be achieved from different initial conditions and in different ways (Gresov and Drazin 1997). Equifinality also requires a guiding vision and simple rules invoked during decision-making (Davis et al. 2009; Eisenhardt and Piezunka 2011; Sull and Eisenhardt 2012). Nevertheless, despite its emphasis on recursivity, this literature only captures known interactions across two adjacent levels. A finer-grained approach is needed to better understand how the dynamics of coordinated search over hidden interactions are nested across multiple and non-adjacent levels.

Time and organizational change

Time – as duration, rate of change, frequency, delays, timing or sequence – has a central place in the study of complex systems and in organization science (Kaplan and Orlikowski 2013). A multi-level approach requires the timescales of observation to be congruent with the levels at which an organizational phenomenon is investigated. For example, the pace of change in a department might necessitate observations over a period of days or weeks, while changes occurring at the level of an organization as a whole could require observation over months or quarters.

All complex social systems involve time delays, whereby the consequences of changes may take time to be noticeable. Mechanistic thinking, where cause and effect are seen as close both in space and time, assumes changes implemented within a system result in an observable impact almost immediately within the same system. But this does not account for a system's inertia, which induces

time delays before changes in behavior are noticeable. Moreover, as the number of interdependencies between state variables increases at higher levels of analysis, so does a system's order. This is similar to increasing the time delay between the input and the lagged response. Therefore, a difficulty in managing complex systems is that corrective actions are often based on perceptions of the *current* dynamics, which were themselves induced by *previous* interventions. Further correction is influenced by the perceived discrepancy between the system's targeted and current state and rate of change, without taking into account the cumulative effect of previous changes to which the system is already responding, albeit with inertia (Sterman 2000). Hasty correction may dissipate progress, especially since activating change in a complex system often leads to a period of "worse before better" (Forrester 1971). Consequently, information processing requirements increase with the order of the system because more state variables and additional integrations must be taken into account.

Moreover, organizations can carefully sequence their actions and focus attention on different time scales (Brown and Eisenhardt 1997). Actions can be sequenced to heighten or reduce interdependencies so that the degrees of freedom of agents are reduced by prior choices of action by other agents (Levinthal and Warglien 1999). Similarly, certain decisions are more central than others because they affect a larger number of other decisions (Rivkin and Siggelkow 2007).

Research Design, Data, and Methods

We follow a qualitative approach based on the theoretical selection of empirical cases (Eisenhardt and Graebner 2007) with 1) high organizational complexity, and 2) where the processes of local search were observable across multiple levels. Investigating multiple cases allows us to generate more varied empirical evidence and increase the external validity of our substantive theory. Healthcare provides a suitable context because of its acknowledged complexity, with many agents distributed across multiple levels in highly interdependent sub-systems (Sterman 2006). We were fortunate to have access to an ongoing national change program delivered across multiple local healthcare organizations, a disruption to a healthcare system which from the start challenged agents' initial cognitive representations.

The UK healthcare system

The UK's National Health Service (NHS) oversees the delivery of healthcare, coordinating local organizations responsible for planning and providing services across functional sub-systems such as primary care (e.g. general practitioners, GPs), secondary care (e.g. acute and community hospitals), and social care, ambulance, or mental health services. In Scotland, the location of our research, fourteen "health boards" are responsible at a regional level for coordinating services. The healthcare system has a typical multilevel structure. Focusing on secondary care, we propose the following levels of analysis: level 1 - individuals within a department, level 2 - departments within a hospital, level 3 - acute and community hospitals within secondary care, level 4 - primary, secondary, ambulance and other healthcare services within a regional health board, and level 5 – all regional health boards within the Scottish NHS system. Interdependencies occur across boundaries at one level (e.g. teams within a hospital department), across adjacent levels (e.g. different departments within a hospital), or with more distant parts of the broader system (e.g. another hospital in secondary care, GPs in primary care).

The Unscheduled Care Collaborative Programme

In August 2004, Scotland introduced a policy to reduce waiting times in accidents and emergency (A&E) and improve the quality of unscheduled care. A performance target for hospitals to assess, treat and admit or discharge 98% of patients arriving in the emergency room within four hours was introduced. A national initiative, the Unscheduled Care Collaborative Programme (UCCP), was launched in May 2005 to help health boards meet the target by December 2007. A 98% target represents a benchmark for world-class performance in unscheduled care, hitherto rarely achieved as it requires a highly robust system, able to absorb great variability in attendances.

The UCCP comprised a national team supporting local implementation in each regional health board. Some managers were recruited from the English NHS, where the target had been introduced in 2002, providing opportunities to learn from their experience. The UCCP drew on lean thinking concepts to manage four patient flows – minor injury and illness, acute assessment, medical admissions and surgical admissions. The approach rested on a collaborative methodology developed by the Institute of Healthcare Improvement in the USA to highlight the interdependencies between

departments and processes typical of an acute hospital. Local teams addressed each patient flow, with flow leaders and program managers nominated at several system levels (national team, regional health boards and each acute hospital). The UCCP introduced a "plan-do-study-act" (PDSA) model of short-term and incremental experiments to help agents discover the nature of their interdependencies and monitor the impact of the changes on performance. From inception, the rhetoric of policy documents, UCCP toolkits and management presentations emphasized "complexity," system interdependencies, the need for "whole-system" change, and the empowerment of agents for their local search:

> *"The target is a recognized measure of whole-system design and capacity. This means all elements of the service in hospitals and in the community are involved in meeting the target – it is not just accident and emergency departments. Engagement across whole health and care systems is needed to make the necessary improvements – all parties are encouraged to think about the way the whole service delivery system works, rather than focusing only upon their own service." NHS Scotland, 2006 policy document (our emphasis)*

> *"Change will not be delivered by issuing one-size-fit-all directives. Solutions must meet local needs and circumstances and more importantly actively engage staff in the change process if significant and sustainable improvement is to be achieved." NHS Scotland, 2005 policy document*

Unlike previous unscheduled care improvement programs focusing solely on the hospital's "front door" (level 1), the UCCP developed priority rules and guidelines to concentrate local teams' search efforts towards better synchronization across hospital departments (level 2) *and* with "out-of-hospital" services (level 3, 4, 5). The national UCCP leaders knew that the English program had not paid enough attention to the interdependencies with GPs, ambulances, or social care, so a fifth patient flow was introduced to help agents discover the nature of and collaboratively address these higher-order interactions in order to divert patients from attending the emergency room or speed-up discharge.

(1) *"A lot of people, at the beginning, thought this was really just all about how the A&E department worked, rather than how the whole system worked. From GPs through to the back door and discharge […] people at local level, who didn't work in A&E and hospitals, often kind of thought 'what has this got to do with me?'" National team member.*

The national team advised local teams, monitored progress towards the four-hour target, and organized national events where participants explained the interdependencies they had discovered, shared the results of their local experiments, and recognized that they were all confronting similar issues.

Data collection

Our research question required a multilevel approach (Klein and Kozlowski 2000) using a multiple embedded cases methodology (Yin 2014). In the context of a single national change program, we studied how individuals conducted changes within departments, across departments in the major hospitals, across local care systems in four different health boards, and across national healthcare organizations. We gathered primary data (exploratory interviews, a questionnaire to select embedded cases, in-depth semi-directed interviews, observation notes and photographs) and secondary data (policy and UCCP documents, and performance data at hospitals, health boards, and national levels).

The first phase of data collection took place between December 2006 and June 2008, covering in real time the final twelve months of the UCCP and a period of six months after its official completion. This helped to mitigate retrospective bias by allowing us to focus both on events within the program during its earlier phases and on its real-time evolution (Leonard-Barton 1990). We initially conducted exploratory discussions with the national UCCP team, allowing us to verify the whole-system approach that the program was trying to achieve. We then organized a workshop with eight representatives from six regional health boards to discuss their experiences and the challenges they faced in meeting the target. These participants answered a questionnaire with closed and open-ended questions about the changes induced by the UCCP in their health boards. The workshop and questionnaires gave us an initial insight into how these agents perceived the interdependencies in their local health systems. We read the background documents produced by the Scottish Government, NHS Scotland, and the national team, such as wider policy documents, the rules for structuring the patient flows, toolkits and guidelines for local search, and internal UCCP reports on performance. We also attended two national UCCP events where we listened to presentations, recorded field notes, and photographed posters by local teams on their experiments and performance improvement trajectories.

Based on these initial data, we selected four health boards for detailed case research, anonymized here as HB1-4. These were chosen to reflect different demographics, initial performance data and progress towards meeting the target. Following receipt of NHS ethical approval, we collected data on the major acute hospitals located in these health boards, read further policy and program documents, and conducted interviews at health board and hospital levels. We conducted 78 individual semi-

structured interviews with informants from diverse positions, hierarchical levels, organizations, and formal or informal roles in the UCCP. We interviewed all the program managers and most flow leaders at each hospital and health board. Interviewees included nurses, doctors, hospital directors, data managers, health board CEOs, GPs, social workers and national NHS managers. Interviews lasted around one hour and generated over a thousand pages of single-space interview transcripts.

We presented our preliminary findings at a workshop for the Improvement and Support Team of NHS Scotland and the Scottish Government, and at the final UCCP national event attended by a critical audience of about 150 delegates. During a second phase, we conducted further interviews with national leaders in 2013 and 2015, and collected secondary data on subsequent policy changes and on waiting-time performance at the hospitals, health boards and national levels up to December 2014.

Data analysis

Data analysis was conducted in two main stages, following Azoulay et al. (2010). In the first we started by exploring the theoretical dimensions of complexity theory applicable to organizational adaptation: *self-organization*, *joint search*, *time dynamics*, *dissipative processes, system structure,* and *system behavior*. Guided by definitions found in the literature we deconstructed these dimensions into second-order constructs, enabling us to structure our raw data in terms of the theoretical constructs whose relationships we aimed to extend. In the second stage, we used the causal loop diagram method, widely used in system dynamics (Azoulay et al. 2010; Sterman 2000) and organization studies (Weick 1979), to generate an explanation of the multilevel dynamics of organizational adaptation. Our case-study allowed us to get empirically closer to the theoretical constructs, their causal relationships and the underlying change dynamics over time (Siggelkow 2007).

Stage 1: theoretical deduction of the constructs

We looked for evidence of *self-organization* as an outcome of the tension between *top-down imposition* (via *authorization and support*, *guidance*, *hierarchical imposition*, *monitoring* and *targets*), and *bottom-up emergence* (via *empowering local agents* and *equifinality*). We looked for evidence of *joint search* as an outcome of tensions resulting from experimentation *within* and *across* boundaries. Within boundary experimentation relates to the *collecting of data* in order to evaluate *local incremental actions*. These may result from the *diffusion* of practices that are then *adapted* to the local

context. Across boundary experimentation relates to the *cognitive representations* of agents from different organizational parts of their coupling interdependencies; their propensity to engage in *collaboration* and *synchronization* of their local search processes (Gavetti and Levinthal 2000; Knudsen and Srikanth 2014; Levinthal and Warglien 1999). We coded for the multiple dimensions of time - statements that related to *duration*, *frequency*, *sequence*, *delays* or *aperiodic coupling*. We also introduced several codes specific to the UCCP data to finely capture the multilevel structure of Scotland's healthcare system and the possible links between the organization level of interactions and time dynamics. Hence, we refined *frequency* as daily, weekly, monthly, or yearly and coded *level* into the five levels at which processes occurred (e.g. dynamics taking place within a department, hospital or higher level). We coded statements relating to *dissipation* across a boundary via the constructs *inefficiency of efforts*, *active resistance* (e.g. as when informants talked about sabotage) and *erosion.*

During this stage, we maintained an inductive stance to identify important themes emerging from the data. We found that managers engaged in what we call *boundary work* by leveraging the classical constructs of *legitimacy*, *sensegiving* and *rhetoric*. Similarly, we realized that the construct *cognitive representations* had to relate to the mental map that agents have of the structure of interdependencies within which they are embedded, rather than their views of the target.

Stage 2: empirical induction of the causal dynamics

Given our interview data, we can assume within-group homogeneity whereby, at each organizational level, members of an organizational group are "sufficiently similar" on a given construct that they can be considered as a whole (Klein et al. 1994; Klein et al. 1995). We also assume that between-group variability is important. As we will present in the findings, this means, for example, that a hospital is composed of specialist departments that are heterogeneous in their cognitive representations of the "whole-system" problem space, while individuals within each specialist departments are quite homogeneous. These assumptions hold recursively across the multiple levels of our case study. Hence, the levels of analysis in our theory development are defined by the levels of healthcare organization (departments, hospitals, primary care, regional health board, national health system).

Finally, we analyzed data coded in stage 1 to uncover the causal relationships at play between these constructs during the process of organizational adaptation. We captured the positive or negative

polarity of these interactions in a causal loop diagram substantiated by the empirical evidence (Azoulay et al. 2010). We started this analytical process with *cognitive representation* because this construct is central to the theory of organizational adaptation. Our empirical data showed that the differences in cognitive representations of different agents, regarding both the existence and nature of epistatic interactions, impacted on collaboration across boundaries, influencing the time taken for synchronized changes to occur and amount of effort used and dissipated. Hence, we started our causal map with the variable *cognitive distance*. We then looked at which other constructs were impacted by *cognitive distance* by systematically working through data and extracting causal relationships. Some reinforcing (R) and balancing (B) feedback loops emerged among the causal relationships. The resulting causal map provides an internally consistent explanation of the change dynamics observed within and across all organizational levels for each embedded case.

Findings

Across all the cases studied, it was clear that agents' initial cognitive representations of their interdependencies with others played a key role in the dynamics of organizational adaptation. It took time for many agents to understand the "whole-system" approach advocated by the UCCP:

(2) *"Right from the beginning of the programme, we were told that waits in A&E departments are not an A&E problem, they are our whole system's problems. But we had to prove, I suppose, to ourselves, that it wasn't an A&E problem, and in order to do that we had to change the way we worked […] there were a great many people who were very, very skeptical of, one, the need for it, and two, the logistics of actually making it happen. But, I think, people became aware that, actually, things were getting better." (Level 2) FL-AH#14* [1]

Interdependencies and organizational search

By adopting a whole-system approach, agents had to jointly discover their interdependencies with other agents across several levels and to engage in coordinated search with them to design novel ways of working while still achieving their respective local performance target. Table 1 presents an overview of the stakeholders upon whom those changes depended, the nature of the interdependencies, and some of a multitude of process or physical changes that were introduced across the different sites.

Insert Table 1

The national team presented data to clearly demonstrate that, contrary to common belief, attendances to emergency departments are extremely predictable on a daily basis. Within emergency departments (level 1), many attendances are from people with minor conditions who self-present at an acute hospital. One of the first changes introduced across all our sites was a process of "minor streaming" whereby minor injuries or illnesses could be quickly recognized and treated. These patients were sent to dedicated "see and treat" areas staffed by emergency nurse practitioners (ENP) and emergency consultants for more serious cases. This filtering had an important impact during "out-of-hours" periods, when the GPs (level 4 interdependency with primary care) are not open and the general public is more prone to self-present at emergency departments. In some hospitals, these see-and-treat areas were staffed by GPs during those periods following a change in their national contract (level 5).

A variety of changes were introduced within acute hospitals (level 2) to improve the accuracy and speed of information across a patient's journey, to facilitate transfer of patients across departments, and to manage bed capacity more precisely. These actions aimed to improve the downstream discharge process to free inpatient capacity earlier in the day and allow patients to move through the hospital, hence limiting backlog and upstream waits. By sharing more accurate information across departmental boundaries, hospitals were able to calculate an estimated date of discharge (EDD). In all hospitals, introducing an EDD required changes in the practices of many agents, from A&E clerical staff having to use a new information system to surgeons in medical wards having to estimate a patient's date of discharge based on their condition, as well as more regularly and consistently inputting their decisions into the information system. Consultants were asked to change their ward round routines, to start with the most critical patients, but then, instead of moving "clockwise" through the ward, to attend to the patients most likely to be discharged. The EDD information was also used during daily bed management meetings to identify those beds to be freed early in the day, critical for timely hospital admission of both emergency and elective care patients. Indeed, specialists in the different departments still had to fulfill their own waiting time targets for elective care or diagnostics.

At several hospitals new non-medical roles, "bed busters", were created to swiftly clean up and prepare available beds without taking time from nurses. To reduce transport or information delays across departments, some radiology departments were moved closer to A&E, new procedures were

introduced to obtain laboratory results quicker, a new role of "flow coordinator" was created, hospital porter rostering was redesigned and even carpets were removed in one hospital to make new trolleys roll more easily. Some hospitals introduced discharge lounges for patients waiting for transportation, pharmacy prescriptions, or other care arrangements. In one hospital the pharmacy was relocated closer to the discharge lounge to quickly provide patients with their prescription before sending them on their way. However, this type of buffering solution was not recommended by the just-in-time approach of the UCCP, because such buffers did not address the ultimate causes of delay.

Novel ways of working with agents *outside* the hospital (level 3, 4, and 5) also had to be found to address uncovered interdependencies. Some elderly patients require social care packages to be put in places before they can leave the hospital. Many sites introduced a "traffic lights" system displayed in the ward. Using the EDD and clinical data, newly created discharge planning teams categorized patients and identified those ready for discharge but still waiting for transportation or social care. In some cases, social care workers were initially concerned that such traffic lights would be used for "*focusing blame*". Similarly, when elderly patients required ambulance transportation there was some concern from ambulance services that their emergency response time targets could be impacted.

The initial emphasis of the UCCP was on streamlining the discharge process and working backwards across departmental boundaries, starting from the "back door" of the hospital and the interdependencies with ambulances, community hospitals or social care (level 4). But it was also essential to reduce unnecessary attendances to emergency departments in the first place by coordinating with out-of-hospital agents. Engagement with GPs was most important, partly to ensure that referrals of patients to A&E were timed as far as possible to ensure a bed would be ready if needed. Parallel to the UCCP, GPs across the UK were opting-out of delivering out-of-hours services under a new contract with the NHS (level 5). The delivery of out-of-hours primary care was therefore becoming increasingly reliant on other organizations such as the NHS24 telephone helpline, the ambulance services and emergency departments. Flow 5 (out-of-hospital) leaders worked with NHS24 and GPs to prevent unnecessary attendances at A&E, for example placing GPs in the "see and treat" areas to act as the first point of triage. Collaborative experiments to discover and address level 4 interdependencies included community paramedic schemes with ambulance services, schemes with

pharmacists to provide oxygen concentrators out-of-hours, local directories for care services, and regional campaigns to "educate" the general public about when to attend A&E.

Other (level 5) interdependencies which trickled down involved the coordination of national healthcare policies (e.g. elective care, diagnostics), frequencies of political and budgetary cycles, and the unfolding political agendas of different public sector bodies. Relevant here were the 23 "health, efficiency, access, and treatment" national targets for healthcare delivery, changes in the rotational training of junior doctors, and longer term cycles of public investment in healthcare infrastructure.

Performance trajectories

Cumulatively, the impact of these collaborative changes over time improved performance towards achieving the four-hour target. Figure 1 presents the performance trajectories (percentage of patients discharged from A&E or admitted into hospital within four hours of their arrival against the 98% target) at different levels: a) the main hospital in HB2 with the actions undertaken for organizational search during the initial period August 2006 – February 2007; b) all our case studies of acute hospitals; c) the health board level (all acute hospitals in the health boards to which our case studies belong); d) at the national level (NHS Scotland = 14 health boards).

Insert figure 1

The monthly performance trajectories on figure 1b-d show that, starting from a low performance, the initial improvements were quite rapid but diminishing returns set in and the trajectory became asymptotic to an upper limit. The literature assumes that once agents have reached a performance peak they stop their local search – they plant their flagpole and stay anchored to that peak. However, the empirical data in figure 1 show that the sustainability of peak performance is not assured: performance drifted down over time. What are the processes leading to such performance improvement and decay? We found that the essence of multilevel organizational adaptation is captured by the interplay between three feedback loops: a "boundary work" loop, a "small wins" loop and a "parochialism" loop.

Boundary work feedback loop

Agents from different parts of the system did not necessarily form completely overlapping cognitive representations of the epistatic interactions between them. These external interactions could only be

discovered by broadening their cognitive representations. We adopt the convention that agents *i* were trying to coordinate with agents *j* in another part of the system. As agents *i* engaged with agents *j* to uncover their epistatic interactions, they were confronted with dissonant perceptions. We found that the higher the level at which these interactions occurred, the broader the cognitive representations required to capture them, the greater the initial *cognitive distance* between agents, and the more difficulty agents *i* encountered when trying to convince agents *j* that "it had anything to do with them."

(3) *"It was fairly easy to convince people in A&E that it had something to do with them" (Level 1) PM-AH#12*

(4) *"People intuitively think that if you are working faster, then you end up doing more work. And it was difficult to sell the basic piece of logic that […] the timings are different but the total volume of work is the same. But that again was counterintuitive for people." (Level 2) PM-AH#30*

(5) *"Some clinicians still don't believe it because their perception is that the problem does not lie with them. [For them] the four-hour target is an A&E issue, not a downstream issue, they don't see how they could possibly affect it." (Level 2) FL-AH#8*

(6) *"Trying to engage people who work in that separate service [psychiatry] in discussions about what should happen in A&E has been really very difficult" (Level 4) NT#76*

When confronted by resistance from agents *j,* who dismissed the existence of hidden epistatic interactions between their organizational parts, agents *i* engaged in *boundary work*. This relied on simultaneous rhetorical, sense-giving and legitimating strategies. Informants explained that they used vocabulary emphasizing interdependence and the need for collaboration. These rhetorical efforts mitigated the resistance of agents *j* who viewed the change program only as a managerial target imposed by the government to improve the performance solely of emergency departments. This defensive view allowed agents *j* to refuse to experiment with changes. In turn, agents *i* updated their discourses and talked about "quality of care for the patient", arguing that waiting more than four hours on a trolley in the corridor of an emergency department could not be considered good quality care, no matter how effective the medical treatment eventually received. These rhetorical efforts aimed to create empathy among agents through the mutual recognition of their interdependence (Hogg et al. 2012), but not to create a superordinate identity. Making agents aware that hidden epistatic interactions existed was a crucial aspect of boundary work, as one of the national team leaders told us:

(7) *"I see a project manager as somebody who is actually getting out there and engaging people, and taking on board things that we were trying to promote through the programme, like complex adaptive theory, having simple rules for engagement, using improvement methodology. […] We very much were talking about quality of care and quality of experience, not about targets…because we recognized that we didn't want it to be seen as a management target and a government target." NT#75*

Clinicians across departments often refuted the validity of the generic quantitative data which were initially presented to demonstrate the existence of interdependencies among departments. These clinicians often argued either that the problem was not as prevalent in their own department or that they were a special case, so that the problem did not apply to them. Informants explained how they then learned to adapt their sense-giving efforts by changing the nature and presentation of data to make agents *j* embrace the existence of dissonant organizational interdependencies.

(8) *"With the nursing staff, all you needed to do was pitch them a patient's story and they were sold. We showed [same story] to medical consultants, oh, they went absolutely nuts: 'you can't possibly share someone else's patient and, of course, that was a single patient, and we don't believe your data'." (Level 2) PM-AH#9*

(9) *"Speaking to consultants you adopted a very different style […] than we did to all the other groups of staff. We plotted their performance on an electrocardiograph-type paper, we had a heart and the four flow groups in the four chambers of the heart, so we tried very, very hard to go out of our way." (Level 2) PM-HB#49*

Many informants explained that this boundary work also involved conveying information about the existence of interdependencies via people who would be viewed as legitimate by the targeted agents *j*:

(10) *"Nick brought up this physician […] to talk to our physicians and to say 'look, this is what the four-hour target can do for you' […] because before that it was mostly managers, either myself, or the general manager, or even people from the national team who would be viewed as managers talking to clinicians about how they should change their practice to achieve this target." (Level 2) PM-AH#59*

(11) *"The Flow 5 leader was usually a GP […] That was good because, you know, unless you get some GP's buy-in in the community service, it ain't going to work because doctors will sabotage it. It's what we do and do it very well [laugh]…" (Level 4) NT#72*

Through their rhetorical, sense-giving and legitimating efforts, agents *i* eventually increased the cogency of evidence for agents *j* about epistatic interactions. On receiving sufficient cogent evidence, some agents *j* updated their cognitive representations, decreasing cognitive distance. Yet shifts in cognitive representations were neither instantaneous nor unanimous across agents at any level.

(12) *"It was probably a couple of months before people started to believe that EDD, minor injuries, etc. actually work and began to push that a bit more." (Level 2) DM-AH#31*

(13) *"Some of them are just local GPs that do the unscheduled care at night. Some of them, whom we know, will help you out. But some of the old school GPs who haven't worked in hospitals for years, they won't help you". (Level 4) no role-AH#65*

We found that the accumulation of cogent evidence to broaden the cognitive representation of agents *j* could be dissipated. The lack of continuity of engagement impeded momentum building and seriously dissipated the managerial efforts expended through this boundary-work loop. The training rotation of junior doctors, changing hospital every few months, also briefly, but repeatedly, dissipated previous improvement work. One of the flow 5 leaders at health board level indicated that dissipation around level 4 interdependencies prevented them gaining any traction in their boundary work.

(14) *"We put it down to new junior doctors […] The new doctors just didn't know how we worked. And we found it was very difficult to convince junior doctors about the importance of early discharge planning. […] In August, that just ground to a halt, the new junior doctors […] didn't understand that writing the discharge letter [early] would have a major effect on the number of beds." (Level1) FL-AH#21*

(15) *"We clearly convened a Flow 5 group and we got various people working in the community, social work, nursing, general practitioners, Community Health Partnerships. Some people came and didn't ever come again and then different people came and didn't ever come again. People from NHS24, who agreed, have all suddenly been replaced. Every time I seemed to be speaking to someone new: we never built up a relationship with them." (Level 4) FL-HB#3*

This boundary work loop occurred recursively at each level. Whether agents *i* tried to convince other agents *j* from the same department (level 1) or from another healthcare organization in their health board (level 4), they needed to engage in boundary work to present convincing evidence and close the cognitive distance between them. However, the cognitive distance increases with the level at which their interdependencies occur. Convincing more distant agents of the existence of epistatic interactions requires more boundary work, which in turn gets increasingly dissipated at a higher level.

Small-wins feedback loop

In the small wins feedback loop, as cognitive distance with some agents decreases, synchronized search with collaborative PDSA initiatives across organizational boundaries begins (see table 1). In turn, mutual performance improves and small-wins at every level are reported and increase the cogency of the evidence presented to further reduce cognitive distance.

(16) *"We got the enthusiastic people who wanted to make it work, who could see the possibilities. Pilots were put in place, and that won the hearts and minds of others who saw it working. Once we started to get improvements and the results all began to move up, people began to understand what we were talking about, and it was easier and easier to persuade other people. And then, once they started to see a real outcome in terms of their own patient flows, I think they realized it was a good idea. We had the proof of this." no role-HB#25*

Agents *j* were initially unwilling to engage in collaborative search across organizational boundaries because they rejected the external epistatic interactions. Cognitive distance meant that agents *j* often refused to change their behavior by experimenting with new ways of working. Given their initial cognitive representations, they inferred that these experiments would be beneficial solely to agents *i* and detrimental to their own target *j*, even if coordination would have improved their own performance; an "illusory hill climb"(Levinthal and Warglien 1999). Agents *j* maintained a parochial focus on attaining their own target, without considering the externalities whose evidence they had dismissed and which were not included within their initial cognitive representations:

(17) *"Some junior doctors quite openly at times will say to you 'I'm not interested in your target'." (Level 1) no role-AH#51*

(18) *"In some places, some departments have their own priorities. Whether they're target-related or not, they've got their own priorities. And they don't like having what they see as their priorities disrupted by another department's priorities." (Level 2) NT#77*

(19) *"It was difficult to achieve buy-in from the general [or] orthopedic surgeons, who say 'it's not our target, it's your target [...] The ambulance services were saying, 'we're a national organization and we've got our targets, so we appreciate your targets, but our targets are more important, whatever happens we have to deliver our targets'" (Level 2, 4) FL-AH#54*

Once the cognitive distance between agents *i* and some agents *j* decreased through the boundary work loop, these agents *j* accepted collaboration based on an updated cognitive representation, recognizing the existence of interdependencies whose precise nature had to be discovered. Closing the cognitive distance between them did not eliminate their focus on their own targets. They expected either an improvement in their own performance with at least no deterioration in the performance of agents *i*, or an improvement in the performance of agents *i* with at least no degradation in theirs. We call this a Pareto collaboration. However, such cognitive shifts were not uniformly distributed among agents *j*:

(20) *"Not everybody had bought into it. I think if we had managed to get everybody mentally at the stage that they're at just now right away you'd obviously have seen a massive jump. But it didn't.*

It took us a while to convince people, A that it was important, B that it was really their responsibility to be involved […] People didn't recognize that." no role-AH#45

Once some agents *j* were willing to engage in Pareto collaboration with agents *i*, they tried to synchronize their local search efforts. Nevertheless, it was not immediately evident in which part of their respective problem spaces they should pursue their experiential search. As table 1 shows, synchronization across boundaries was not about common incentives, but rather about the principles of coordination and information exchange. Such synchronization was guided by two mechanisms: the national team guidance for sequencing search towards particular "long-jumps" and vicarious learning.

(21) *"[national team] was absolutely critical, because it said, 'whoa, hang on a moment, you've all got these ideas, you're all rushing off doing things, but these are the things that we actually think you need'. Some of those things weren't necessarily our priorities […] They said 'that should be your priority'. It helped focus." (Level 2) no role-AH#22*

Vicarious learning occurred during events organized by the national team where agents from across the entire Scottish healthcare system could share learning about successful actions to redesign interdependencies across departments or more distant organizational parts. Through cognitive experimentation based on their representation of their own problem space, other agents inferred whether these actions would work for them too, and if so to adapt them to their local circumstances:

(22) *"Definitely...we used the national workshops… people talk about what they were doing […] how they had implemented it and what their outcomes had been. [In some cases] we thought: 'that sounds as if that might be a possibility'. We then contacted a few other health boards to see if they were doing a similar thing and they were, but they all had slightly different models […] We kind of devised our own to meet our own immediate needs and adapted that. I would say it was mostly from looking at what was happening elsewhere." (Level 5) PM-AH#5*

Nevertheless, the evidence for these mutual improvements could only be obtained after three time delays. First, a change would take time to implement. Second, the system had inertia which, according to many informants, increased with the level at which the epistatic interactions occurred:

(23) *"I mean that [level 1 change] was just amazing, just an amazing piece of work, so simple, so effective, I'd never seen a change so dramatic overnight, and I mean overnight, literally." (Level 1) PM-AH#5*

(24) *"Flow Five is very definitely a longer-term piece of work. It [level 4 change] must have taken me about three, four months to do it." (Level 4) FL-HB#16*

Third, the frequency of the monitoring determined the time required to obtain data and present evidence. The national team advised which data to collect, but there were often delays in recruiting

data managers, setting up the monitoring system, and collecting enough data to present cogent evidence to local agents. As the frequency of monitoring decreased inversely with level, the reporting time delay increased at higher levels. The total time lag on this part of the small wins loop is the sum of implementation time, system inertia and reporting delay, all of which increased at higher levels:

(25) *"I think it was very rapid. Once we started making these changes, you could look at charts and see when the changes were made, and what difference it made." (Level 1) FL-AH#29*

(26) *"I reckon the first year, at least the first fifteen months… I don't think we really started getting people focusing on it until this year because then the data was coming back and that really focuses you: when you see your name and your position on a list." (Level 2) NT#74*

(27) *"Once improvements have been tested and shown to be effective and sustainable, you may decide to stop measuring temporary measures monthly and decide to measure them quarterly. You will need to have five separate data points that show sustainable improvement before you can reduce the frequency of measurement." UCCP national program toolkit*

Given the accuracy of their representations, agents could not precisely infer ex-ante the impact of a change on performance; it was part of the UCCP program methodology to try out incremental changes and monitor their impacts. If improvements could be shown, the change was maintained and the evidence used to refine agents' understanding of their problem space. However, as many incremental changes were tried throughout a healthcare organization, it was sometimes impossible to link an action to an outcome. Nevertheless, whatever the causal link, what mattered was that performance improved:

(28) *"When we plotted all the changes we'd made, the performance was steadily improving, although we couldn't demonstrate by using statistical process control which individual change had been the cause of that. People could see that they had contributed to this major improvement: that was a very good positive reinforcer." (Level 2) PM-AH#12*

Like in the boundary work loop, the lack of continuity in engagement was a source of dissipation within this small wins loop. For example, the training of junior doctors (a level 5 policy) required them to rotate among hospitals, often inducing periodic disruptions. Crucially, kick-starting the small wins loop relied on the sustained collaboration of a few early-initiates:

(29) *"It relied on enthusiastic individuals and if these enthusiastic individuals were on holiday or off for any reason, the performance really just kind of fell and we were slightly alarmed at how easily things went back to just the way they were." (Level 4) FL-HB#4*

Parochialism feedback loop

As described by several informants, when performance reached the 98% target diminishing returns rendered further progress difficult and the local search plateaued (figure 1):

(30) "At the beginning you were tackling huge chunks that made a big difference, but once you get into the nitty-gritty, when they were at 97, 98 percent consistently, it's the last 3 percent that's the hardest because it's the hardest thing to change." (Level 2) no role-AH#28

(31) *"We don't really have any ideas left as to what we would additionally do… that last 5 percent is proving very, very difficult indeed. There are many factors that are quite difficult to fix: there are physical factors in terms of the design of hospitals." (Level 2) FL-HB#47*

Another reason for slower progress is that given the performance trajectory achieved, local agents across all the organizations often lifted their foot off the pedal.

(32) "We're coming to a bit of a barrier, people's ideas are drying up […] they say they've got 'change fatigue' and I'm asking more and more of them all the time." (Level 5) NT#71

(33) *"The downside is, as we reached the peak actually. There's still a way to go and people, knowing that they've been successful, are now just prepared to sit back. […] An unexpected outcome is people reveling in a glory that's not there yet, because there has been such improvement. I think that could come back and bite." (Level 4) PM-HB#32*

Such diminishing returns rendered coordination of search harder and pushed agents *j,* outside A&E, to revert to a rather myopic focus on their parochial targets rather than to keep accounting for the higher-order interdependencies which had been uncovered and jointly redesigned. Targets across organizational boundaries competed for scarce attention and Pareto collaboration decreased:

(34) "I remember a consultant radiologist saying that she was getting shouted at because the four-hour target was going to be breached; because the 62-day cancer target was a problem for a particular patient; and she had to do a radiology to hit the inpatient target. So, you know, they've all got that competing demand." (Level 2 and 5) PM-HB#58

Erosion of performance

During 2005-2008, sources of dissipation countered performance improvements, reducing the traction gained via the small wins feedback loop: lack of continuity of personnel (different people attending meetings, rotation of junior doctors), active resistance, and causal ambiguity (lack of clarity about the impact of individual actions, delays due to system inertia or frequency of monitoring). Moreover, changes occurred at different health system levels and with different frequencies, creating an unstable

environment which dissipated ongoing progress: new national policies (level 5), strategic reviews of health board services (level 4), or physical changes to infrastructure in some hospitals (level 2).

(35) *"… the NHS was going through tremendous change: national modernizing took three years, the changes with the junior doctor training. We've been trying to cope with a lot of other changes and within [HB1] we've had the review of services […] which has coincided with the timing of this [UCCP]." (Level 4) FL-HB#4*

When asked about the sustainability of changes, the consensus was that since the new way of working was so much better, it should become embedded. Some informants however identified that progress could be eroded after the UCCP formally ended because the focus on interdependence could be lost:

(36) *"… if we disband with all the flow groups my worry would be that as new staff come on board … the stories which are relayed to them will be different […] if we're not doing that sharing, the bonding, the teams coming together. It's about not losing what we've learnt … [It's about] sustaining those networks with your colleagues across, continuing to share information with them, not becoming complacent […] I am worried that the whole system philosophy will be lost if we don't have the single system teamwork that we've got through working roundtables with multidisciplinary staff" FL-AH#7*

Since 2008, after the UCCP ended, there have been cyclical dips in performance, like in social dilemmas where cascades of competitive behaviors are followed by a resurgence of collaboration back to the cooperative equilibrium (Levinthal and Warglien 1999). However, the cyclical dips in figure 1d have increasing amplitude, especially during "winter crises", and decreasing recovery. These may be due to the conflating effects of several dissipative mechanisms. In June 2010 the mandatory 98% 'target' was converted to a "minimum standard" and lowered to 95%, with weaker sanctions on CEOs. During the following year's winter crisis performance significantly deteriorated. Staff turnover, including the departure of UCCP managers, also led to increased cognitive distance and a loss of organizational memory about unsuccessful experiments:

(37) *"People forget what it takes to be a 98% organization, a very robust and stable organization that can absorb volatility. People who understood moved on, and new people arrive… Organizations … are converting back to things that were proven not to work back in 2006, to … false solutions that are just moving problems to somewhere else" (Level 5) NT#71*

The frequency of monitoring fell, as some changes were believed to be firmly embedded in practices (27), which made deviations less noticeable. Furthermore, more programs were started to meet other targets, described as "*lightning rods*" by a CEO, and managers shifted their scarce attention to the priorities of the moment. Because of diminishing returns, change in incentives and new targets, Pareto

collaboration decreased; because of organizational memory loss, cognitive distance increased and search synchronization fell. The outcome of ceasing to account for interdependencies is an inescapable erosion of performance.

Scale-invariant dynamics

Figure 2 shows how the multilevel dynamics of organizational adaptation in this healthcare system arise from the causal structure demonstrated by our evidence. The same dynamics of boundary work, small wins and parochialism were at play in the interdependencies between individuals within an emergency department (level 1), across hospital departments (level 2), across secondary care organizations (level 3), across a health board's primary and secondary care systems (level 4), or even across national organizations (level 5). But despite this scale-invariance, the relative strength of each loop was a function of the level at which it occurred. Identifying and redesigning higher-order epistatic interactions across higher-level boundaries meant overcoming greater initial cognitive distance between agents, intensifying the boundary-work loop. Initial joint search was slower to establish, greater inertia slowed discovery of the impact of changes, and the time lag in the monitoring of performance increased. The reinforcing (R) feedback loop of small-wins, stronger than the other balancing (B) loops at lower levels, became increasingly weak. At higher levels efforts to synchronize local search across organizational boundaries were increasingly dissipated. While much coordinated search was realized at level 1 (departments) and level 2 (hospitals) throughout the national system, collaborative adaptation at level 3 and level 4 (health board) proved very difficult to achieve.

Insert Figure 2

Discussion and Conclusions

Our findings contribute to the development of theory of multilevel organizational adaptation in three ways. First, we uncover assumptions in previous work regarding cognitive representations and joint search. Second, we highlight the contingency of search on the level of analysis and four aspects of multilevel organizational adaptation which future research should better account for. Third, we present some implications of our results for a scale invariant organization science.

Cognitive representations and joint search

The literature on organizational adaptation emphasizes the importance of self-organization and local action. Individual agents evaluate alternatives based on their current cognitive representations of their environment and their inference of the consequences of their adaptive actions (Gavetti and Levinthal 2000). Existing research often assumes that cognitive representations of interdependent agents are commensurate and reciprocal, can shift instantaneously and that long-jumps are random. Our UCCP case demonstrates that such assumptions are unhelpful in understanding the multilevel dynamics induced by different cognitive representations of agents across organizational boundaries.

Commensurability. The literature is unequivocal on the inaccuracy of agents' cognitive representations; their mental models are of a lower dimensionality than their true problem space. Moreover, given agents' bounded rationality, the inaccuracy of their cognitive representations also increases with problem space dimensionality. Existing work on coordinated search assumes that the mental models of agents across organizational boundaries are commensurate because these agents mutually recognize their external epistatic interactions. The literature in fact mostly considers cases of "epistemic" interdependence relating to the accurate prediction by one actor of another actor's actions. Such epistemic uncertainty is inherent in choices along known dimensions in the respective problem spaces, whose external interactions are mutually recognized by both agents. In the example used by Knudsen and Srikanth (2014), the windmill specialists both know that their choices of shape and material are interdependent; their cognitive representations, even if inaccurate, are commensurate.

While prior work demonstrates that the incongruence of mental models increases with the partitions of knowledge structures, the assumption that the agents' mental models are commensurate remains. Our findings show, however, that achieving coordinated search initially faces the problem of cognitive distance between inter-organizational agents, preventing them from mutually recognizing their interdependence in the first place and independently of the granularity of their knowledge partition on a given dimension. Our empirical findings partly resonate with Puranam and Swamy's (2016) simulations of mutual adjustment processes when agents do not know ex ante how their choices are interdependent, but also demonstrate that they must first recognize their interdependence. Considering organizations as near decomposable hierarchical systems implies that at higher level of

analysis interdependencies are of a higher order and, as our findings show, harder to capture since they require agents to coarsen the scale of their cognitive representations.

PROPOSITION 1: *The cognitive distance between the imperfect cognitive representations of agents from across organizational boundaries, with regards to their external epistatic interactions, increases with the level at which these epistatic interactions are situated.*

Synchronization. Prior work assumes that agents are reciprocally aware they are engaged in a coupled search process, even if they hold asymmetric initial representations of the nature of interdependence (Puranam and Swamy 2016). Thus, agents either expand efforts to align their mental models as their respective local search efforts increasingly partition their individual knowledge structures along known dimensions (Knudsen and Srikanth 2014) or reciprocally integrate firm-wide incentives in their search efforts (Siggelkow and Rivkin 2005). Yet, a key organizational design task is, first of all, to make self-interested agents across an organization, with competing goals, reciprocally aware of their interdependence. This is a precondition for synchronized search and for achieving illusionary hill-climbing (Levinthal and Warglien 1999).

PROPOSITION 2.a: *The synchronization of local search across organizational boundaries requires closing the cognitive distance between agents' imperfect cognitive representations of their external epistatic interactions.*

Boundary work. Experiential learning, and conflict and persuasion within an organization, may challenge agents' cognitive representations of their problem space (Gavetti and Levinthal 2000). Closing the cognitive distance requires some agents *i* to engage in purposive action (Tracey et al. 2011) in order to focus mindsets on collaboration (Liberman et al. 2004) and convince other agents *j* to broaden their cognitive representations by including mutually overlooked epistatic interactions. Cognitive distance, however, leads to dissonance and the rejection of presented evidence (Festinger 1962). When facing such defensive resistance by agents *j,* agents *i* will engage in boundary work to increase the cogency of evidence and induce double-loop learning in agents *j* (Argyris and Schön 1978). Our results show that the interplay between the boundary-work and the small-wins feedback loops offers opportunities to take advantage of the plasticity of cognitive representations in order to

overcome organizational inertia (Gavetti 2012; Tripsas and Gavetti 2000), although the amount of boundary work required increases with cognitive distance.

PROPOSITION 2.b: *Given proposition 1, the boundary work efforts required to close the cognitive distance between local agents' cognitive representations of their external epistatic interactions increase with the level at which these epistatic interactions are situated.*

Instantaneity. Current work on coordinated search is often based on two agents at a single level of analysis and does not account for the time it takes to align mental models. Shifts in cognitive representation are implicitly assumed to be instantaneous as agents immediately discover the consequences of their local search efforts. However, we show that an "updating delay" is at the core of organizational adaptation. First, it takes time for boundary work efforts to close the cognitive distance between agents. Second, all agents from across an organizational boundary will not simultaneously update their cognitive representations based on presented evidence that an epistatic interaction may exist. We assume an average within-group homogeneity (Klein et al. 1994), but our results nonetheless show that some agents updated their mental models earlier than their group peers and were the first to engage in Pareto collaboration in their search efforts, kick-starting the small wins loop whose evidence convinced other agents. This process is not instantaneous but probably reflects the heterogeneous distribution among agents of the strength of prior beliefs (Puranam and Swamy 2016) and, consequently, the amount of evidence required to close a given cognitive distance. The small wins loop has a time delay in the alignment of cognitive representations which increases with scale. Higher-order interactions among multiple agents across multiple levels inhibit instant feedback.

Multilevel adaptation

Our study allows a multilevel analysis with more than the two adjacent levels (individual-organization or organization-context) usually found in the literature. We extend the theory of organizational adaptation by explaining the multilevel nature of its underlying constructs such as search radius, coordination and communication , and feedback and equifinality. We also identify how organizational forgetting and other sources of dissipation affect the sustainability of changes.

Search radius. By coarsening the scale of observation, the ontological nature of components changes and new epistatic interactions of a higher-order are identified. By zooming in to a lower scale,

and given the near-decomposability of hierarchical systems (Simon 1962), epistatic interactions that were considered internal to one system become external epistatic interactions between its sub-systems. Conversely, by zooming out to a coarser scale, interdependencies considered as external epistatic interactions among sub-systems become internal to the meta-system. In our embedded cases, the interdependencies considered as external epistatic interactions between departments at the hospital level become internal epistatic interactions within a hospital when the level of analysis is at the wider health-board level. What is considered as joint search at one level thus becomes local search at a higher level of analysis. Similarly, local search at one level becomes coordinated search at a lower level. This is not accounted for in existing work on organizational adaptation, which usually only considers the impact of the search radius on the speed and extent of performance improvements (Rivkin and Siggelkow 2007; Siggelkow and Rivkin 2005). Our multilevel approach reminds us that at higher levels of analysis, more things change simultaneously and that all these changes combine into a broader search radius than if observed at a lower level.

Coordination and communication. The need for communication increases with the number of epistatic interactions between agents' problem spaces and with the need for coordination (Gavetti and Levinthal 2000). Through constrained communication (Puranam and Swamy 2016) agents coordinate their choices, observe the impact of their mutual actions, and align their respective mental models. The alignment of mental models leads agents to concentrate on a narrow portion of their joint problem space (Knudsen and Srikanth 2014). Such "joint myopia" neglects exploration. Agents are often assumed to explore their problem space via long-jumps which are either random or based on the forward evaluation of attractiveness (Gavetti and Levinthal 2000; Knudsen and Levinthal 2007; Knudsen and Srikanth 2014; Levinthal and Warglien 1999; Rivkin and Siggelkow 2007; Siggelkow and Rivkin 2005). Hierarchy can also be considered as a coordination device, which imposes identical mental models on agents by providing them with the partitions on the dimensions of the problem space (Gavetti 2005; Knudsen and Srikanth 2014).

Our findings identify another mechanism, "guided long-jumps". Extending prior research, we found that hierarchy, which does not necessarily know the local idiosyncrasies and therefore cannot impose a single mental model on all agents, can nonetheless indicate to agents that they should

coordinate around overlooked epistatic interactions in a certain sequence (21). To reuse the windmill example (Knudsen and Srikanth 2014), top-down coordination via hierarchy would not provide agents with similar partitions along the dimensions of materials and shape, but first guide them to coordinate around those two dimensions. The UCCP national team recommended "guided long-jumps" based on simple rules (e.g. "enable early discharge"), as when they indicated to agents in hospitals that they should prioritize the streaming of minor injuries or implement EDD (figure 1a). Such guidance helps agents understand the need to increase the dimensionality of their initial representations, making them aware of interdependencies whose precise nature they should discover and redesign collaboratively.

Our results also highlight the importance of vicarious learning at multiple levels in generating this form of higher knowledge. Scotland learnt from England's program (whose methodology came from the USA) and then adapted it "locally". Following the initial guided long-jumps, agents then learnt what others had done to redesign epistatic interactions and engaged in coordinated search to adapt those solutions to their own contexts. Prior work on organizational adaptation follows the behavioral assumption of intelligent search whereby agents, while lacking omniscience, are capable of identifying proximate alternatives with higher payoff (Levinthal 1997; Levinthal and Warglien 1999). Our findings on vicarious learning challenge the hypothesis of intelligent search, but substantiate the hypothesis of blind experiential local search based on crude cognition (Gavetti and Levinthal 2000).

Feedback and equifinality: Prior work focuses on the coordination of two agents at a single level of analysis and often assumes clear feedback between action and outcome for single-loop learning. Our findings show that the need for coordination increases at higher levels as the interdependencies are of a higher-order and greater system inertia renders the action-outcome linkages more ambiguous. Moreover, the frequency of monitoring decreases at higher levels, increasing time lags among coupled choices. We found there can be causal ambiguity about which actions, among all those within the search radius at that level of analysis, had what impact on performance. The search radius and the time delays increase at higher levels and impede single-loop learning, challenging assumptions of "rich and unambiguous feedback" (Levinthal and Warglien 1999, p. 346) as agents cannot learn from feedback to update their mental models of action-outcome linkages. Such ambiguity in feedback promotes mutual confusion and flawed mental models of the problem space (Knudsen and Srikanth 2014). Our

results show the scale-dependency of this effect and explain why action-outcome linkages are more blurred and harder to interpret as managers move up in the hierarchy (Gavetti 2005).

In fact, our multilevel analysis shows that causal ambiguity does not really matter for organizational adaptation. Following guided-long jumps, experiential search efforts can cumulatively improve performance despite the causal ambiguity of a broad search radius, heterogeneous local solutions and inaccurate cognitive representations. Yet such equifinality raises the question of the sustainability of those improvements. Without a clear understanding of the combined contributions of changes at multiple levels, false positives and negatives impede opportunities to learn.

Organizational forgetting. Organizational adaptation requires both individual and organizational learning (Greenwood and Hinings 1996). Our results demonstrate the key role of agents' cognitive representations in multilevel organizational adaptation but also explain the processes through which they are updated by discovering overlooked epistatic interactions. Organizational learning raises the issue of organizational forgetting and memory loss. Prior work on organizational search assumes that agents remember unsuccessful local experiments (Gavetti and Levinthal 2000). Our analysis indicates that this may not be the case. As postulated by Greenwood and Hinings (1996), our embedded cases show that staff turnover, combined with the successive emphases on multiple targets, leads organizations to forget newly acquired knowledge (Martin de Holan and Phillips 2004) about epistatic interactions and which configurations were successful. Contrary to formal models of organizational adaptation, which assume the sustainability of changes, figure 1 shows that organizational forgetting leads to the significant erosion of the performance improvements achieved.

Dissipation. Treating organizational adaptation as an ongoing process, rather than as a punctuated epiphenomenon, refocuses our attention on the micro-processes of change (Tsoukas and Chia 2002) and requires us to investigate their rates at multiple levels (McKelvey 1997). Our multilevel approach based on complexity theory highlights the importance of accounting for sources of dissipation which both counteract adaptation efforts (e.g. dissipation within the recursive small wins loops prevents reaching a critical mass at each level) and lead to the erosion of performance in the long run (e.g. aperiodic changes across levels, new policies leading to new targets and programs, staff turnover). The

positive scale-dependency of the strength of dissipation is an important result for the theory of multilevel organizational adaptation since it determines which feedback loop dominates in figure 2.

Scale invariance in organization science

There has been a recent call for scale-free theories to reorient the organization science paradigm. Andriani and McKelvey (2009) argue that as the structure of most systems subjected to change is scale-free, scale-free theories are more relevant to the study of non-linear organizational phenomena (Greenwood and Hinings 1996). Although the literature is still vague about the conditions under which organizations exhibit scale invariance, some authors point to several underlying dynamics: path-dependent evolutionary processes (Andriani and McKelvey 2009), cross-level influences (Eoyang 2011) and positive feedback (Morel and Ramanujam 1999). Our analysis strongly indicates that multilevel organizational adaptation depends on the recursive interaction of the same feedback loops across multiple levels. Figure 2 is scale-invariant.

It is usual in the investigation of a complex system at a single level to observe that at each point in time one of the feedback loops will dominate the others and that such dominance will shift over time among the loops (Sterman 2000). Our analysis shows that shifts in feedback loop dominance also occur across levels. At lower levels the small-wins loop can more easily reach a critical mass and induce change within an organizational unit (e.g. department), but as the level of analysis increases and the epistatic interactions are of a higher-order (e.g. level 5) the small-wins loop encounters stronger dissipation, rendering it unable to counteract the parochialism loop becoming dominant.

PROPOSITION 3: *While the feedback structure of an organizational phenomenon may be scale-invariant, the dominance among the interacting feedback loops can be scale-dependent.*

Implications for future research

By increasing the dimensionality of their cognitive representations agents accept synchronizing search across boundaries at multiple levels. For formal modeling approaches, often based on NK(CS) models, future work should clearly specify the level of analysis because, as our findings demonstrate, it has implications for the definition of the search radius and for what constitute K internal or C external epistatic interactions. Moreover, we assume averaged within-group homogeneity at a given level of

analysis, but the data nonetheless show that agents neither align their mental models simultaneously nor instantaneously. A theory of multilevel organizational adaption, based on the assumption of within-group heterogeneity (Klein et al. 1994; Klein et al. 1995; Klein et al. 1999), could therefore be developed to uncover the characteristics of those early collaborators and target them to reinforce the small wins loop. Since cognitive shifts are not instantaneous and there are several time-delays between search actions and observable outcomes, which increase at higher levels, future formal models should also include these time delays in the learning processes occurring during joint-search. More broadly, we call for future theoretical models to truly capture the multilevel nature of organizational adaptation and to better understand the dynamics induced by interdependencies and competing local targets.

[1] For each quote we provide the (Level) of the interdependencies described and the informant's role in the UCCP (no role, FL=flow leader, DM=data manager, PM=program manager, NT=national team), position in the healthcare system (AH= an acute hospital, HB=a regional health board), and informant number #. E.g. informant FL-AH#14 was a UCCP flow leader in an acute hospital, PM-HB#49 was a program manager in a health board.

TABLES AND FIGURES

	Level 1 - within department		Level 2 - within acute hospital		Level 3 - within secondary care		Level 4 - wihtin healthboard		Level 5 - national level	
Stakeholders	Reception staff, A&E managers, A&E consultants, A&E nursing staff		Medical / Surgery clinicians, senior nursing staff, radiography, bed managers, ward managers, pharmacy, porters, facilities managers, hospital managers, IT department, catering, house-keeping, etc.		Discharge coordinator, social care services, community hospitals, psychatric services, famillies, etc.		GPs in primary care, voluntary organizations, ambulance services, social work, community health partnerships, politicians, general public, Patient Focus and Public Involvment (PFPI) groups, health board managers, pharmacists, etc.		NHS 24, Ambulance services, Social Care, Psychiatric, Scotisch Executive, etc.	
	Inter-dependencies	**Solutions**	**Inter-dependencies**	**Solutions**	**Inter-dependencies**	**Solutions**	**Inter-dependencies**	**Solutions**	**Inter-dependencies**	**Solutions**
Information / Processes / Infrastructure	Receive patients Evaluate and Treat Move patients	•Emergency Nurse Practitioners •GP in A&E •Clinical decision unit •See & Treat •Match staff profile to forecasted attendance •Rapid diagnostic technology •Paediatrics streaming •Data collection, clerical input in A&E •Dedicated area for minor injuries •Redesign A&E area •Centralize emergency receiving for surgery / medicine •Increased area for assessment	Admission to wards Treatment and information Discharging patients Capacity Layout Transfer	•Change order of ward round •'Bed Busters' •Estimated Date of Discharge •Flow coordinators •Centralized all planned intermediate and minor surgery •Statistical Process Control •Reduce laboratory results wait •Dedicated radiology for A&E •Redesign porters's time allocation •New beds, new trolleys •Remove carpet •Discharge Lounge •Relocate wards	Discharging patient Involvment of partners in community care Coordinate care Ensure transfer	•Discharge team •'Traffic Light' system •Social Care packages •Coordination with community hospitals	Alternatives to hospital attendances / admissions Reduce waits and delays for patients who need urgent hospital assessment and treatment Timely discharge Involvment of partners in community care Coordinate care Involve in A&E Ensure transfer Planned care	•Out of Hours - GP primary care •GP and other SHO in A&E •Coordination with Nursing Homes / care of the elderly •Review of services •Increasing public awareness of services •Alternatives to paediatrics admissions •Listing of pharmacy, and oxygen tanks •OT/Physio in A&E •Support Service Partnering •GP medical referrals •Mental Health Service emergency in rural settings •Acute and homeless liaison	National Targets	•Targets and standards •Health and Social Care Directorates •UCCP National Team •National data

Table 1 : Multiple levels, stakeholders, interdependencies, and examples of solutions during organizational adaptation

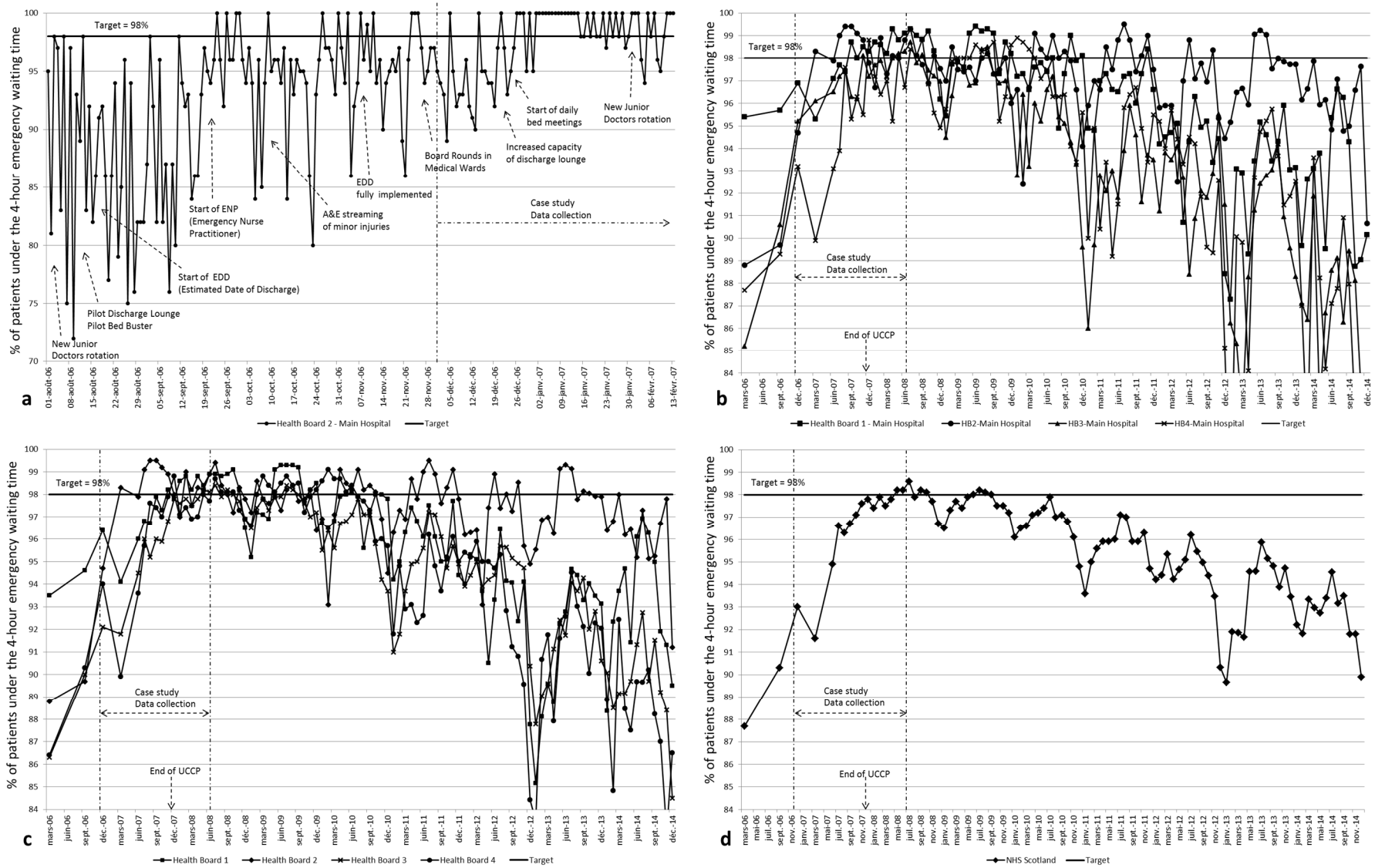

*Figure 1: 2006-2014 "4-hour emergency waiting time" performance trajectory: **a**. main hospital of health board #2 with sequence of interventions, **b**. main hospitals in case study health boards, **c**. all hospitals in case study health boards, **d**. across all health boards (14) at national level*

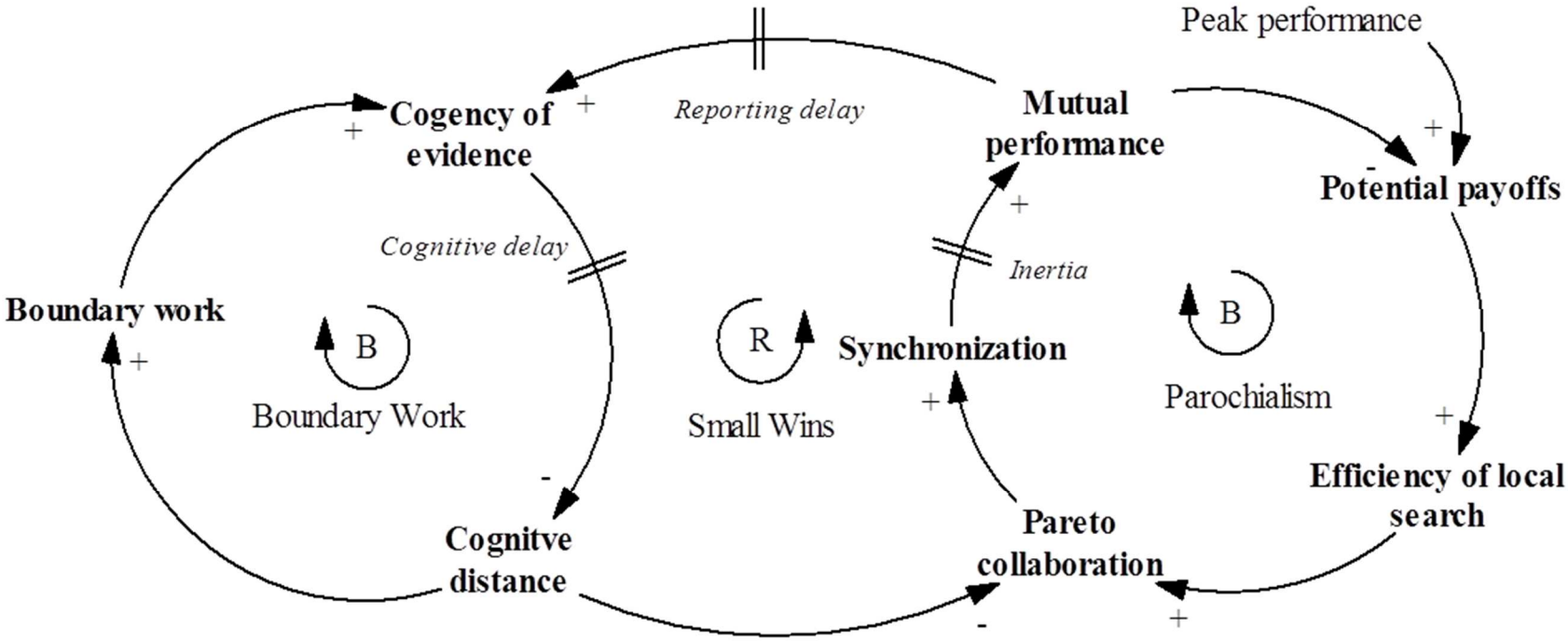

Figure 2: Scale-invariant dynamics of multilevel organizational adaptation (R reinforcing, B balancing, feedback loops)

REFERENCES

Allen, P., S. Maguire, B. McKelvey, eds. 2011. *SAGE handbook of complexity and management*. SAGE, London.

Anderson, P. 1999. Complexity theory and organization science. *Organization Science* **10**(3) 216-232.

Anderson, P., A. Meyer, K. Eisenhardt, K. Carley, A. Pettigrew. 1999. Applications of Complexity Theory to Organization Science. *Organization Science* **10**(3) 233-236.

Andriani, P., B. McKelvey. 2009. From Gaussian to Paretian Thinking: Causes and Implications of Power Laws in Organizations. *Organization Science* **20**(6) 1053-1071.

Argyris, C., D. Schön. 1978. *Organizational learning: A theory of action perspective*. Addison-Wesley Publishing Company.

Azoulay, P., N.P. Repenning, E.W. Zuckerman. 2010. Nasty, brutish, and short: embeddedness failure in the pharmaceutical industry. *Administrative Science Quarterly* **55**(3) 472-507.

Boisot, M., J. Child. 1999. Organizations as adaptive systems in complex environments: The case of China. *Organization Science* **10**(3) 237-252.

Brown, S., K. Eisenhardt. 1997. The art of continuous change: linking complexity theory and time-paced evolution in relentlessly shifting organizations. *Administrative Science Quarterly* **42**(1) 1-34.

Davis, J., K. Eisenhardt, C. Bingham. 2009. Optimal structure, market dynamism, and the strategy of simple rules. *Administrative Science Quarterly* **54**(3) 413-452.

Eisenhardt, K., H. Piezunka. 2011. Complexity theory and corporate strategy. P. Allen, S. Maguire, B. McKelvey, eds. *SAGE Handbook of Complexity and Management*. SAGE, London, 506-523.

Eisenhardt, K.M., M. Graebner. 2007. Theory building from cases: opportunities and challenges. *Academy of Management Journal* **50**(1) 25-35.

Eoyang, G. 2011. Complexity and the dynamics of organizational change. P. Allen, S. Maguire, B. McKelvey, eds. *SAGE Handbook of Complexity and Management*. SAGE, London, 317-332.

Festinger, L. 1962. *A Theory of Cognitive Dissonance*. Stanford University Press.

Forrester, J.W. 1971. Counterintuitive behavior of social systems. *Theory and Decision* **2**(2) 109-140.

Gavetti, G. 2005. Cognition and Hierarchy: Rethinking the Microfoundations of Capabilities' Development. *Organization Science* **16**(6) 599-617.

Gavetti, G. 2012. Toward a Behavioral Theory of Strategy. *Organization Science* **23**(1) 267-285.

Gavetti, G., D. Levinthal. 2000. Looking Forward and Looking Backward: Cognitive and Experiential Search. *Administrative Science Quarterly* **45**(1) 113-137.

Greenwood, R., C.R. Hinings. 1996. Understanding radical organizational change: Bringing together the old and the new institutionalism. *Academy of management review* **21**(4) 1022-1054.

Gresov, C., R. Drazin. 1997. Equifinality: Functional Equivalence in Organization Design. *Academy of Management Review* **22**(2) 403-428.

Hogg, M.A., D. van Knippenberg, D.E.I. Rast. 2012. Intergroup Leadership in Organizations: Leading Across Group and Organizational Boundaries. *Academy of Management Review* **37**(2) 232-255.

Kaplan, S., W.J. Orlikowski. 2013. Temporal Work in Strategy Making. *Organization Science* **24**(4) 965-995.

Kauffman, S., S. Levin. 1987. Towards a general theory of adaptive walks on rugged landscapes. *Journal of Theoretical Biology* **128**(1) 11-45.

Klein, K.J., F. Dansereau, R.I. Hall. 1994. Levels issues in theory development, data collection, and analysis. *Academy of Management Review* **19**(2) 195-229.

Klein, K.J., F. Dansereau, R.J. Hall, W.S. Sherman, R. Grant, Y. Fried. 1995. On the level: Homogeneity, independence, heterogeneity, and interactions in organizational theory. *Academy of Management Review* **20**(1) 7-17.

Klein, K.J., S.W. Kozlowski. 2000. *Multilevel theory, research, and methods in organizations: Foundations, extensions, and new directions*. Jossey-Bass.

Klein, K.J., H. Tosi, A.A. Cannella Jr. 1999. Multilevel theory building: benefits, barriers, and new developments. *Academy of Management Review* **24**(2) 248-248.

Knudsen, T., D.A. Levinthal. 2007. Two faces of search: Alternative generation and alternative evaluation. *Organization Science* **18**(1) 39-54.

Knudsen, T., K. Srikanth. 2014. Coordinated Exploration: Organizing Joint Search by Multiple Specialists to Overcome Mutual Confusion and Joint Myopia. *Administrative Science Quarterly* **59**(3) 409-441.

Leonard-Barton, D. 1990. A dual methodology for case studies: synergistic use of a longitudinal single site with replicated multiple sites. *Organization science* **1**(3) 248-266.

Levinthal, D.A. 1997. Adaptation on Rugged Landscapes. *Management Science* **43**(7) 934-950.

Levinthal, D.A., M. Warglien. 1999. Landscape Design: Designing for Local Action in Complex Worlds. *Organization Science* **10**(3) 342-357.

Lewin, A.Y., H.W. Volberda. 1999. Prolegomena on coevolution: A framework for research on strategy and new organizational forms. *Organization science* **10**(5) 519-534.

Liberman, V., S.M. Samuels, L. Ross. 2004. The name of the game: Predictive power of reputations versus situational labels in determining prisoner's dilemma game moves. *Personality and social psychology bulletin* **30**(9) 1175-1185.

Maguire, S., B. McKelvey, L. Mirabeau, N. Ötzas. 2006. Complexity science and organization studies. S.R. Clegg, C. Hardy, T.B. Lawrence, W.R. Nord, eds. *The SAGE Handbook of Organization Studies*, 2nd ed. Sage, London, pp. 165–214.

Martin de Holan, P., N. Phillips. 2004. Remembrance of Things Past? The Dynamics of Organizational Forgetting. *Management Science* **50**(11) 1603-1613.

McKelvey, B. 1997. Quasi-natural organization science. *Organization Science* 352-380.

McKelvey, B. 1999. Avoiding Complexity Catastrophe in Coevolutionary Pockets: Strategies for Rugged Landscapes. *Organization Science* **10**(3) 294-321.

Moldoveanu, M.C., R.M. Bauer. 2004. On the Relationship Between Organizational Complexity and Organizational Structuration. *Organization Science* **15**(1) 98-118.

Morel, B., R. Ramanujam. 1999. Through the Looking Glass of Complexity: The Dynamics of Organizations as Adaptive and Evolving Systems. *Organization Science* **10**(3) 278-293.

Puranam, P., M. Swamy. 2016. How initial representations shape coupled learning processes. *Organization Science* **27**(2) 323-335.

Rivkin, J.W., N. Siggelkow. 2007. Patterned Interactions in Complex Systems: Implications for Exploration. *Management Science* **53**(7) 1068-1085.

Siggelkow, N. 2007. Persuasion with case studies. *Academy of Management Journal* **50**(1) 20-24.

Siggelkow, N., J.W. Rivkin. 2005. Speed and Search: Designing Organizations for Turbulence and Complexity. *Organization Science* **16**(2) 101-122.

Simon, H.A. 1962. The architecture of complexity. *Proceedings of the American Philosophical Society* **106**(6) 467-482.

Sterman, J.D. 2000. *Business Dynamics: System Thinking and Modeling for a Complex World*. Irwin McGraw-Hill, New York.

Sterman, J.D. 2006. Learning from evidence in a complex world. *American Journal of Public Health* **96**(March (3)) 505–514.

Sull, D., K. Eisenhardt. 2012. Simple Rules for a Complex World. *Harvard Business Review* **90**(9) 68-74.

Thietart, A.-R., B. Forgues. 2011. Complexity science and organization. P. Allen, S. Maguire, B. McKelvey, eds. *The SAGE Handbook of Complexity and Management*. SAGE, London, 53-64.

Tracey, P., N. Phillips, O. Jarvis. 2011. Bridging institutional entrepreneurship and the creation of new organizational forms: A multilevel model. *Organization Science* **22**(1) 60-80.

Tripsas, M., G. Gavetti. 2000. Capabilities, cognition, and inertia: Evidence from digital imaging. *Strategic Management Journal* **21** 1147-1161.

Tsoukas, H., R. Chia. 2002. On Organizational Becoming: Rethinking Organizational Change. *Organization Science* **13**(5) 567-582.

Weick, K.E. 1979. *The Social Psychology of Organizing*. Random House, Incorporated.

Weick, K.E. 1995. *Sensemaking in organizations*. Sage Publications, London.

Yin, R.K. 2014. *Case study research: Design and methods*, 5th edition ed. SAGE.

Zhou, Y.M. 2013. Designing for Complexity: Using Divisions and Hierarchy to Manage Complex Tasks. *Organization Science* **24**(2) 339-355.