EXPERIMENTATION IN ORGANIZATIONS: AN INTEGRATIVE REVIEW

Leonardo Corbo

Bozidar Vlacic

Riitta Katila

November 2025

EXPERIMENTATION IN ORGANIZATIONS: AN INTEGRATIVE REVIEW

Abstract

Organizations have long relied on experiments to guide decision-making. Yet, a comprehensive synthesis of this rich and timely empirical literature remains lacking. In this integrative review, we identify two primary streams of research that shape our understanding of experimentation in organizations: (1) problem-solving–based experimentation and (2) causal-inference–driven experimentation. The problem-solving stream emphasizes iterative experimentation, learning from failure, and navigating organizational challenges, while the causal inference stream focuses on sharp identification, structured experimental designs, and bounded experiments, each rooted in distinct disciplinary traditions. Despite the differences, these perspectives offer complementary insights into how organizations experiment, learn, and adapt. Through an analysis of 177 empirical studies across several disciplines, and integrating these parallel streams, we develop a unifying framework that highlights key drivers, processes, and outcomes of experimentation in organizational contexts. We conclude by outlining promising avenues for future research, including deeper retention in shaping experimental effectiveness and organizational learning, overcoming cognitive biases, expanding the scope of experiments to strategy, organizational design and people processes, and the possibility of the two streams to cross-feed: problem-solving to generate broad hypotheses that causal inference sharpens, and causal inference experiments to trigger reframing of the problem-solving experiments.

In this integrative review, we take stock of research on experiments *by* organizations *within* organizations. While organizations have long used experiments to resolve key uncertainties in decision-making (e.g., Staw, 1977) in a range of functions from product development to manufacturing and marketing (March, 1981; Huber, 1991), there is no comprehensive synthesis and integration of this rich and timely empirical literature. With the interest in causal experiments (RCTs, A/B tests, switchbacks, etc.) surging, the literature and our understanding of experimentation in organizations remains fragmented, and no integrative review has yet to compare the variety of approaches– a gap that we address in this review. We particularly ask how experimentation contributes to organizational outcomes such as innovation or learning, and how our understanding of experimentation in organizations has evolved over time.

The term "experiment" comes from the Latin word experimentum, a noun of action built on the verb experīrī, "to try" or "to test." This origin is aligned with the literature we reviewed: experiments are fundamentally about "trying" something and generating evidence through direct experience. Organizations try something to learn more, such as to determine the efficiency of something previously untried. The notion emphasizes the active nature of the process, the inherent uncertainty (most experiments fail), and the result of learning something new, even if through failure. Consistent with the literature that we reviewed, we define *experimentation in organizations* as testing or trying something in an organization through direct experience with the result of learning something new or more about it.

An integrative review of experimentation is timely for several reasons. Since early calls to evaluate public policy programs (Suchman, 1968) and early interest in organizational experimentation (e.g., Staw, 1977; Huber, 1991), research on experimentation in organizations has ebbed and flowed over the years (see Figure 1).[1] Today, there is a resurgence of interest in

[1] While our core emphasis is experiments by organizations, the genesis of experimentation traces back to calls for experiments in the 1950s to systematically investigate the effectiveness of social intervention programs (e.g., education). Campbell (1969) made the case that policy and program decisions should emerge from continual social experimentation that tests ways to improve social conditions, which could at the extreme lead to developing

experimentation, driven by the availability of big data and the prevalence of platform organizations that offer ample opportunities for testing: thousands or more experimental units can be tested with relative ease, allowing organizations to effectively track changes and outcomes (Kohavi et al., 2020; Larsen et al., 2024; Reineke, Katila, & Eisenhardt, 2025). This shift in real-world organizations coincides with scholarly interest to advance robust causal inference and the renaissance of interest in the scientific method and evidence-based movements in organizations in general.

Our detailed analysis of the literature reveals two distinct streams of organizational experimentation.[2] The first, which we term problem-solving-based experimentation (P-S), traces its roots to operations and engineering research and emphasizes iterative experimentation, learning from failure, and organizational complexity, including identity, power, and politics. P-S studies conceptualize experimentation as a trial-and-error process in which each trial generates new insights and cumulatively helps narrow the scope of future attempts (Allen, 1966; Rosenberg, 1982; Sitkin, 1992; Thomke, 1998).[3] Yet, in organizational practice, implementation and learning from experiments may often emerge as uncoordinated, boundedly rational, and political. Empirical designs in this stream include for example prototypes and MVPs. Altogether, in the P-S view, experiments are iterative tests for workable solutions that prioritize relevance and feasibility over isolating a single causal ingredient.

The second stream, which we term causal-inference-based experimentation (C-I), conceptualizes experimentation as the process of designing an intervention and systematically

an "experimenting society" (e.g., Campbell, 1969). While this work (often labeled as program evaluation) focuses on federal-government and state level policy and our focus, in contrast, is organizations, subsequent work on experimentation built on this early policy evaluation work.

[2] Policy evaluation studies are similarly split in two: (a) holistic evaluations: contextual evaluation models which holistically examine particular program operations, and (b) focused evaluations: experimental evaluation models which simplify program realities in generalizable analyses of discrete causes and effects (Britan, 1978; Rossi, Freeman, & Lipsey, 1999). At a very high level, these streams correspond to problem-solving vs. causal-inference streams that we have outlined in this review.

[3] In these studies, the word experiment is sometimes used interchangeably, but somewhat imprecisely, with the word experience (i.e., an entirely different construct). Miner et al. (2003) is an example: it is often cited as a study of organizational experimentation, but most of the paper documents effects of experience, not experiments.

observing the outcomes in order to establish a causal relationship (Cook & Campbell, 1979; Fisher, 1935). Rooted in the philosophy of science (Hacking, 1983), the causal inference stream portrays experimentation as structured, and systematic. Implementation and precision of the intervention may, however, vary based on the actual organization design elements, and many experiments remain isolated one-offs. Thus, building a cumulative, generalizable knowledge base is a challenge. Empirical designs in this stream include for example A/B testing and switchbacks. Altogether, in the C-I view, experiments are tightly-scoped tests designed to estimate unbiased causal effects of well-specified but often narrow interventions.

Altogether, our review identifies and traces the historical evolution of the two parallel but loosely connected streams on problem-solving and causal inference. We outline the distinct yet complementary disciplinary roots (operations and engineering research versus philosophy of science), and diverging emphases of process versus performance in each stream. We also identify each stream's distinct constructs, theoretical frameworks, design choices, and empirical approaches. Comparing and contrasting the streams offers an opportunity to identify overlapping approaches, emergence of a common "language" to explain experimentation, and research gaps that can be better understood in the future. A deeper understanding of experimentation in organizations also aligns with the broader goals of the management field to improve causal identification strategies (Thau, Pitesa, & Pillutla, 2014; Hauser et al., 2017).

Integration of problem-solving and causal-inference streams can help scholars provide better evidence-based knowledge to practitioners as well. By clarifying the core concepts, juxtaposing different perspectives and unifying the scattered evidence, our integrative review can help scholars generate richer research agendas and produce insights that generalize across contexts, and, ultimately, help practitioners deploy experimentation more effectively.

The review is also timely, as despite the surging interest, systematic integrative reviews on experimentation are scarce. Existing reviews focus on particular applications of experiments, such as A/B testing, but lack broad and general coverage of organizational experimentation and clear inclusion criteria. Other reviews examine related constructs, such as discovery-driven planning, lean startup, probe and learn, or purposeful experimentation (e.g., McGrath, 2024; Shepherd & Gruber, 2021). However, these studies are more focused contributions and typically conceptual, not reviews of empirical work. Other research (e.g., Thau et al., 2014) examines experimentation as a research method for organizational scholars, but does not examine experimentation *by* organizations, which sets those studies outside the scope of the current review. Our review thus fills an important gap. The rapid rise of organizational experimentation—driven by digitalization, A/B testing, RCTs, and business strategy applications (e.g., lean startup)—provides an opportunity to systematically compare the historical literature with recent work. Some researchers have even dismissed the recent excitement as "old wine in a new bottle," suggesting it would be useful to identify commonalities and disagreements over time in the experimentation literature, further emphasizing the need to take stock of this rapidly expanding body of work.

There are several contributions. First, we identify two streams of experimentation – problem-solving and causal inference – and clarify how they can complement and strengthen one another. We organize the literature in both streams using the classic variation–selection–retention (VSR) framework (Campbell, 1969), which Aldrich (1999) highlights as a foundational lens for understanding organizational change processes, including experimentation. We also identify the types of threats to validity (internal, external, construct, statistical conclusion) that are most salient for each experimental stream.

Second, by reviewing the research across management and adjacent fields, we bring together perspectives from radically different areas, emphasizing the common ground and the differences across the approaches. Overall, our review highlights several persistent challenges in the literature: (1) a lack of cumulative knowledge-building across research communities;

particularly between recent empirical work and earlier foundational studies; (2) while many studies claim to examine experiments as intentional actions only, they often lack precision in defining intentionality, and, in practice, many experiments are emergent or unplanned; (3) despite decades of theorizing, no organization appears to have fully realized the "experimenting organization" ideal across all functions (Staw, 1977), first envisioned in the 1950s.

Several boundary conditions, and so opportunities to expand through future work, emerge through our review. While learning is often discussed as one of the key objectives for experimentation (Huber, 1991; Katila & Ahuja, 2002), our VSR analysis reveals an intriguing open question around *retention gaps*, i.e. whether organizations actually learn. In the short term, the question is whether the organization commits to acting on the experiment's results, and in the long-term whether this learning is retained over time – both open questions that are currently addressed by neither problem-solving nor causal-inference work. Greater focus on retention as the missing link that explains why experimentation does not always yield learning is a key finding. A related challenge for organizations is the content of learning: most improvements from experiments are incremental and isolated. As firms gain experience, experiments, particularly causal inference ones, emphasize refinement of, and attention to ever-narrower issues (Thompson, 2010). This pattern is akin to the "Twitter effect" on search landscapes where attention clusters around the same issues and drifts away from potentially new and different topics (Bahceci et al., 2015) as well as from the holistic view of what the organization is trying to accomplish through its experimentation program. How experiments might instead serve as a source of organizational change and new direction remains an open question that is rarely discussed in either stream.

Second, our review indicates that most empirical studies focus on *cognitive reorientations* (i.e., pivots) in the post-experimentation phase (Berends et al., 2021; Grimes, 2018). Despite the importance of cognition for organizational action (Helfat & Peteraf, 2015), only a few studies focus on mindset shifts and identity dynamics *during* experiments in organizations, forming interesting avenues for future work.

Third, the *scope* of experimentation provides avenues for future work. While many experiments in the studies we reviewed are run to tweak product features or pricing, experimentation in fundamental organizational domains such as governance (e.g., organizational structure and design choices; Reineke et al., 2025) or corporate strategy is rare. Testing new structures, new team processes or strategy alternatives is difficult also due to the significant disruption it can cause to the organization. Currently there is no clear way (digital twins and DAOs may be an exception; Reineke & Katila, 2025) to conduct low-risk, low-cost causal inference experiments in these domains that would fulfill validity criteria.

The rest of the paper proceeds as follows. In Section 2, we explain the methodology we followed to identify and analyze the relevant literature. Section 3 presents an overview of the literature, illustrating the drivers, processes, and consequences of experimentation, while emphasizing differences and commonalities between the two streams. Section 4 discusses the key insights from our exploration, the integrative conceptual framework, the practical implications for organizations, and the open questions for future research.

[INSERT FIGURE 1 ABOUT HERE]

METHODS

In this section, we explain the scope of our review, the rationale behind the selection of the papers, and the process we followed to analyze them. To best represent the current state of the literature and formulate specific recommendations, we performed multiple searches using Web of Science, Scopus, and ABI/Inform research databases (Rousseau, 2024).

To identify relevant empirical articles on experimentation in organizations in management and adjacent fields, we initially conducted searches using a Boolean search string to retrieve manuscripts whose titles, abstracts, or keywords directly targeted experimentation in organizations. Specifically, we compiled publications by searching the following series of keywords: “business* experiment*”, “discovery-driven growth”, “discovery-driven planning”, “experimental learn*”, “experiment* in organization*”, “lean start-up*”, “organization* experiment*”, “purposeful experiment*,” or “probe and learn”. This initial search yielded a total of 580 articles from Web of Science, 1,016 from Scopus, and 564 from ABI/Inform. An overview of our sampling protocol is presented in Figure 2.

[INSERT FIGURE 2 ABOUT HERE]

The initial step involved combining the retrieved manuscripts, resulting in a total of 2,160 publications. Next, we removed the duplicates (n=903). Subsequently, we manually re-screened the titles, abstracts, and author keywords of the remaining 1,257 articles to determine their relevance for full-text screening.

In Step 2, we carefully read all articles and excluded those not explicitly focused on experimentation in organizations. This reduced the sample by 959 articles. In Step 3, an analysis was conducted on the remaining 298 manuscripts. An additional 37 articles without a clearly defined focus on experimentation in organizations were excluded, resulting in a preliminary sample of 261 manuscripts, which consisted of 147 empirical studies and 114 non-empirical articles.

Given the rich literature and growing interest in experiments within organizational settings, we established additional explicit boundary conditions. Peer-reviewed publications written in English were the primary focus, while relevant non-peer-reviewed sources were also reviewed to provide a thorough account of the topic’s history and development (Rousseau, 2024). During the stepwise sampling process, we specifically targeted studies on experimentation in organizations. We excluded research in fields such as medicine or physics, classroom-based studies (e.g., education), studies on enabling technologies such as AI, and papers that only peripherally referenced organizational experimentation.

In step 4, we expanded our search to include studies examining well-known design choices, including A/B tests, Beta tests, Minimum Viable Products (MVP), pre/post assessments, prototypes, switchbacks, and Randomized Controlled Trials (RCTs). Our in-depth article review and consultations with experts in the field guided our use of snowball sampling to identify additional relevant studies. Consequently, we employed backward (reference-based) and forward (citation-based) snowballing (Larsen, 2002). Among the identified papers, we specifically searched for those that, in addition to initial keywords, explicitly addressed the aforementioned experimental types. Moreover, consultations with experts enabled us to identify important studies that were not captured through either the initial search or snowballing procedures. This expanded search yielded 6,127 articles from backward snowballing and 10,522 from forward snowballing.

In Step 4a, we merged these snowballing results, resulting in 16,649 manuscripts. Step 4b involved removing duplicates and previously screened-out manuscripts, reducing the total by 5,799. Following another screening of titles, abstracts, and author keywords (steps 4c and 4d), we excluded an additional 97 manuscripts after an in-depth, careful full-text review due to a lack of explicit focus on our topic. This process resulted in a final snowball sample of 43 articles, including 30 empirical and 13 non-empirical studies.

Finally, we merged preliminary (n=261) and snowball samples (n=43) to form our general sample of 304 articles. Within this general sample, 177 were empirical studies, constituting the core empirical foundation for our subsequent analysis and synthesis. The list of publications is available in Appendix A.

Following Glynn and D'Aunno (2023) and Reineke et al. (2025), we analyzed the 177 empirical studies to generate insights about experimentation in organizations. We first coded papers based on the P-S (67% of the papers) and C-I (19% of the papers) logics of experimentation. When studies followed both logics, we coded them as both P-S and C-I (14%

of the papers). We then identified the key drivers of experimentation by looking at individual, team, organizational, and external aspects as well as the consequences, which were mainly related to innovation, knowledge acquisition or dissemination, growth and performance, and adaptation and change. Based on the variation, selection, and retention (VSR) processes underlying experimentation (Aldrich, 1999; Campbell, 1969), we also classified the organizational experiments illustrated in the sampled papers by sources of intentional variation, selection and elimination of certain types of variations by intervention and retention. Per Aldrich, intentional variation occurs when individuals actively generate alternatives and seek solutions (e.g., formal experimentation programs, incentivized employee experiments, gamification, or playfulness like Google's 20% rule). Selection involves eliminating certain variations (e.g., external vs. internal, competitive pressures, or outdated criteria), while retention reflects learning. The main methods employed by our reviewed studies were qualitative (55%) and quantitative (35%), and 10% used mixed methods.

We also identified the primary empirical contexts and geographic locations in which experimentation in organizations was conducted. The majority of the studies looked into experimentation within organizations located in major economies, including China, Italy, and the United States, while several studies adopted multi-country approaches to trace experimentation across different organizational locations and cultural settings. Regarding industry contexts, the technology sector and digital platforms are predominant, followed by more traditional industries such as automotive and finance. Numerous studies embraced a multi-industry perspective. Next, we evidenced the interdisciplinary character of experimentation in organizations (see Table 1), as evidenced by its distribution across management (41%), entrepreneurship and innovation (33%), computer science, operations and engineering (18%), and other fields (8%). This breadth highlights the available evidence across various disciplines as well as potential opportunities for integrative, interdisciplinary insights.

[INSERT TABLE 1 ABOUT HERE]

EXPERIMENTATION IN ORGANIZATIONS: PROBLEM-SOLVING AND CAUSAL INFERENCE STREAMS

Through integrative review, we identify two primary streams of research that shape our understanding of experimentation in organizations: (1) problem-solving-based experimentation and (2) causal-inference-based experimentation. Consistent with the work we reviewed, we organize the literature within each stream using the classic variation–selection–retention (VSR) framework (Campbell, 1969).[4] Murray and Tripsas (2004) argue that "experimentation occurs across firms through a process of variation, selection and retention" (p. 46), and Levinthal (2017) discusses how the Mendelian executive makes decisions by following a process of variation, characterized by problem-framing, planning and intentionality, and selection, where the executive interprets experimental outcomes and makes a conscious selection among alternative possibilities. In a study of medical device executives, Katila et al. (2017) similarly classify organizational roles based on whether the role emphasizes variation or selection, and find that different emphases activate distinct aspects of a professional's expertise, which in turn influence firms' experimentation and retention of innovation outcomes.

Using the VSR framework, we first review and synthesize the work on the antecedents of experiments: how and why *variation* is introduced. Second, we consider interpretation of experiments: *selection* of certain variants while others are discarded. Third, we address institutionalization of selected variants and the conditions under which they are *retained* over time. After analyzing the VSR process in each stream, we discuss general organization-level outcomes of experimentation.

Problem-Solving Stream

[4] Variation is defined as any departure or deviation from current routine or tradition, such as changes in products, routines, organizational forms etc. that are tested and tried. Variations are the raw materials from which the selection processes cull those that are most suitable, and possibly also retained over time.

As noted above, P-S studies frame experimentation as iterative tests. In this stream, experimentation prioritizes relevance and feasibility over isolating a single causal ingredient.

Variation

The first step in experimentation is framing the problem. This means deciding what variation to test. The key idea is that the higher the scope of variation, the greater the opportunities for change through experiments. In practice, variation arises through multiple mechanisms. In the work we reviewed, we categorize variation into intentional versus blind; and into its individual, organizational and external determinants.

Blind variation. Across much of the problem-solving experimentation literature, variation emerges through ad-hoc processes. Austin, Devin, and Sullivan (2012), for example, describe a process of "accidental innovation," where valuable unpredictability fuels creative experimentation. Such blind variation typically comes from unstructured, spontaneous ideation rather than deliberate planning of experiments. A well-known historical example is the Manhattan Project (Gillier & Lenfle, 2019). Variation, born from necessity and urgency, illustrates how ambiguity and even random decisions can drive inventive experimentation. Rather than follow a structured sequence of trials, teams advance somewhat "blindly" by iterating on partial ideas and improvising experiments under conditions of radical uncertainty (c.f., Garud and Karunakaran, 2018).

Decentralization, i.e. distributing decision-making authority to lower levels of the organization often goes hand in hand with variation. In particular, decentralized decision-making tends to foster more blind variation as it allows diverse actors at the local levels to independently generate and test ideas (Bahceci et al., 2015; Reineke et al., 2025). Drawing on evidence from the drone industry, Bremner and Eisenhardt (2021) show that community-based organizational forms—characterized by less hierarchy and greater inclusivity—are especially conducive to generating blind variation. In contrast, traditional hierarchical firm structures often constrain such variation and favor conformity (Reineke et al., 2025).

Sometimes variation is not blind, but just extremely small-scale or marginal (an anomaly). Lego is one example. Thomke and Loveman (2022: 8) document how Knudstorp, the executive chairman and former CEO of the Lego Group described that "even when the percentage of customers complaining about a product is extremely small, a company should really listen and listen very actively." In the problem-solving stream, these kinds of anomalies, no matter how small, are often embraced as interesting variations to be explored further.

Closely related is variation that emerges through deviant behaviors such as rule violations, workarounds, or unintended accidents (Aldrich, 1999). Especially in contexts marked by bureaucratic rigidity (experiments are viewed as too risky) or by resource constraints (experimentation is not cost-effective), such deviance can spark meaningful experimentation. A classic example is 3M's creation of Post-it Notes, which began as a failed attempt to develop a super-strong adhesive and evolved into a product born of accidental discovery, informal experimentation, and meaningful recombination.

In some cases, failed experiments themselves become a source of unexpected variation. Alblas and Notten (2021) document how setbacks in the manufacturing of semiconductor equipment frequently gave rise to non-planned, blind experimentation. When production processes are disrupted—due to inexecutable operations, deviations in expected outcomes, or defective components—engineers initiate "debug iterations," short development loops that test new (unexpected) assumptions or small design changes. Similarly, Khanna, Guler, and Nerkar (2016) observe that in the pharmaceutical industry, high failure rates in early-stage trials could be embraced as a sign of productive variation, particularly in novel therapeutic domains. In both settings, failure is not simply a setback but a generative moment that can prompt useful variation and a new path for experimentation.

Intentional variation. At the other end of the spectrum, a select set of problem-solving studies portrays experimentation as strategic and deliberately structured. Rather than blind, variation – i.e. the problem-framing - is purposefully structured through formal programs of experimentation, leadership behavior, or systematic tools or activities (Burgelman, 1983). Thomke (1998) offered an early example. In product design and innovation, techniques like rapid prototyping and simulation allow organizations to experiment purposefully, generating design alternatives in a cost-effective and structured manner. In this view, variation is not left to chance but embedded intentionally within formal development processes to maximize learning and innovation. Dalpiaz et al.'s (2016) study of the Italian manufacturer of household goods Alessi is another example. The authors describe "several deliberate efforts to combine the institutional logics of industry and art." In this example, hybrid designs are intentionally used to generate novel variation to experiment with. Another example is organizational storytelling that was described as a way to test alternative value propositions, encouraging symbolic variation.

Finally, it should be noted that despite the recent rise of theory-based-experimentation movement (e.g., Camuffo et al., 2024), only a few of the real-world experiments and experimentation programs in organizations that were reviewed used robust theory as a starting point. To wit, Gillier and Lenfle (2019: 450) write: "Manhattan Project experiments exhibit an important lack of theoretical knowledge…and the absence of pre-established organization". Of course, there are exceptions like scientific experiments run by scientists in organizations, but outside science, the use of theory is still rare.

Individual and team factors underlying variation. Within the intentional variation stream, several individual- and team-level factors drive the type of variation that is experimented with. One explanation is (in)experience of individuals that often goes hand-in-hand with experimentation. Studies particularly link decision-makers' seniority (Huber, 1991;

Thomke, 2020), both in age and professional experience, to resistance to experimentation. Na et al. (2022), in a study of women entrepreneurs, find that younger entrepreneurs tend to pursue more experiments than their older counterparts, suggesting that life stage and generational perspectives influence variety in experimentation. Leatherbee and Katila (2020) have a similar finding regarding professional experience. Entrepreneurs with extensive experience of business-oriented training are often more inertial towards business-model experiments whereas those with engineering backgrounds (i.e., limited experience of business) are more likely to embrace variation and experimentation. Bingham and Davis (2012) also suggest that prior experience of executives impacts whether firms engage in experimentation, or rather embrace indirect forms such as imitation. Overall, experience increases inertia and limits variation.

Within small organizations, Solaimani et al. (2022) observed that variation (and the related experimentation) is shaped by the founders' frustrations and vision of the future. Individual attributes like self-efficacy (Chen & Zhang, 2024) and entrepreneurial orientation (Zhao et al., 2021) also matter. Zhao and colleagues (2021) found a relationship between entrepreneurial orientation of senior executives in Chinese companies and experimental learning. They link this finding to the context of transitional economies where organizations are constrained by a more limited access to external resources and up-to-date technological knowledge and so encouraged to experiment. Finally, Kirtley and O'Mahony (2023) find that early-stage entrepreneurs experiment with their strategies only when "obligated", i.e., when new information conflicted with their beliefs. Overall, variation is often dampened by inertia.

In large firms, a key obstacle to variation is risk-aversity (reluctance to adopt an experimental mindset), even in the face of promising innovations (Weissbrod & Bocken, 2017). Risk is often tied to high likelihood of failure inherent in an experiment. Despite recognizing the inherent learning value of failures, organizations frequently exhibit reluctance toward high-risk experimentation that may result in failure. Lee et al. (2004) highlight this paradox, emphasizing that experiments that are more likely to result in failures also provide critical insights and opportunities for long-term innovation through for example mapping the search terrain (Sitkin, 1992), yet organizations and their managers are often reluctant to engage in this type of more "risky" experimentation. Thomke (1998) similarly underscores that, even when economic costs are minimal, the prospect of failure remains a deterrent to experimental variation, often because of career concerns and hesitance to take risks and fail. These findings underscore that failure avoidance may critically shape the scope and direction of variation in organizational experimentation, especially in large established teams and established organizations.

Leadership styles and team-level factors also play a role in shaping variation. For example, Lee, Edmondson, Thomke, and Worline (2004) show that leaders who frequently shift strategic direction can generate substantial variation in project trajectories, thereby encouraging more extensive experimentation. However, the (positive) effect of such variation is contingent on team dynamics: when psychological safety is high and evaluative pressure is low, teams are more willing to explore such multiple experimental paths. In contrast, high evaluative pressure threatens experiments and may render variation ineffective or even harmful (Lee et al., 2004). Beyond shifts in strategic direction, leadership can be relevant simply from the resource allocation perspective. Because experiments can be costly and time-consuming (Loch et al., 2001; Thomke, 1998), leadership support is essential to support variation.

Another important leadership-related factor shaping variation is a startup founder's identity frame (i.e., an individual's internal understanding of "who I am" and "who I want to be" as an entrepreneur). In an inductive study of four early entrants in the emerging air taxi industry, Zuzul and Tripsas (2020) demonstrate how variation within startups is filtered through the founder's self-concept. Some founders embrace experimentation broadly, generating wide-ranging variation in ideas and strategies. Others deliberately constrain variation in order to maintain coherence with their sense of self. In this way, identity acts as a cognitive filter—shaping not only what paths are explored, but also what possibilities are dismissed. Intriguingly, Zuzul and Tripsas (2020) trace the differences in variation back to the depth and nature of founders' prior experience. Firms led by founders with deep technical expertise in a single domain—such as software or aviation—were less likely to experiment whereas those with broader but less specialized backgrounds were more experimentally-minded, and led more variation-generating startups. Crucially, the authors find that the attitude towards variation and experiments is not simply a matter of experience, but also of identity. Garud and Karunakaran's (2018) account of Gmail is another example. Gmail challenged Google's identity since "any efforts that diverged from the core identity around "search" were viewed with suspicion". To avoid the political challenge of selling an identity-challenging invention to the organization, the inventors of Gmail re-purposed existing Google code, used their day-off for experiments, and invited other Google employees to become users of the product.

A rich body of research adds the importance of teams and highlights the value of cross-functional composition. In particular, interdisciplinary composition of teams often promotes divergent thinking and parallel exploration. Gillier and Lenfle (2019) refer to this as heuristic prototyping, i.e., a process where design unfolds based on experience and intuition rather than predefined goals, allowing cross-functional teams to pursue multiple partial solutions simultaneously. Brown and Eisenhardt (1997), in their study of six computer industry firms, similarly, propose that teams composed of members from different functional backgrounds contribute diverse perspectives, which in turn fuel variation in experimentation, and Dalpiaz et al. (2016) demonstrate how Alessi introduced variety into its experimentation through collaborations with designers from a range of creative disciplines, each bringing distinct expressive sensibilities and capabilities.

Finally, some research finds that role-competence also goes hand in hand with variety that feeds experimentation. In particular, studies in medical device product development show that professionals do not perform equally well during all stages of experimentation—some excel in generating variation, while others are more effective in selecting among alternatives (but are less open to trying out new variation in the first place) (Katila et al., 2017). To understand the differences, the authors introduce the idea of role competence. To match the individual with the right experimentation stage, they classify organizational roles based on whether a particular organizational role emphasizes variation, selection, or a combination of both. They theorize—and find empirical support—that different roles activate distinct aspects of a professional's expertise, which in turn influences experimentation, and innovation outcomes. For example, physician–CEO role, which often prioritizes selectivity and risk mitigation, may unintentionally suppress experimental tendencies of physicians and de-prioritize variation (Katila et al., 2017). Altogether, this misalignment can have significant consequences for organization's experimentation and change.

Organization-level factors underlying variation. Other studies examine organizational levers of experimentation and point out that organizational context also plays a pivotal role. The role of organizational architecture and design is particularly important in shaping variation. As noted above, studies make the case that decentralization drives variation (Reineke et al., 2025 for a review). For example, distinguishing between formal and informal structure, Contigiani and Young-Hyman (2022) found that experimentation is positively welcomed by external audiences only when associated with the appropriate structure.

At the ecosystem level, West and Iansiti (2003) highlight how platform leaders such as Microsoft and Intel can catalyze varied experimentation on the platform. By offering tools and formal infrastructure that enable third-party developers to try different approaches, platforms can stimulate a wide range of experiments, effectively inducing variation across the ecosystem.

Expanding on this, Thatchenkery and Katila (2021, 2023) show that the way firms perceive their competitors within an ecosystem, including the platform's power to constrain, or support, the ecosystem's creativity, significantly shapes which types of variation complementors pursue and how broadly they experiment in new directions.

Interorganizational alliances also serve as a fertile ground for generating variation for experiments. In a longitudinal case study, Browder et al. (2022) demonstrate how teams in vendor-client relationships challenge prevailing innovation logics and experiment in cross-sections of interest akin to a "Venn diagram"—they produce novel variations in mutually interesting directions that would likely not emerge within a single organizational boundary. Similarly, Bremner and Eisenhardt (2021) identify distinct experimental trajectories across different types of organizational forms—corporate ventures, independent startups, and portfolio entrepreneurs—each shaped by its own strategic logic and structural constraints. Notably, portfolio entrepreneurs generate more novel and diverse combinations by engaging in boundary-spanning experimentation that draws from multiple domains.

External factors underlying variation. Variation in experimentation can also emerge through iterative market feedback. Lynn et al. (1996) were among the first to label this approach as "probe and learn," describing how firms like Corning, Motorola, and GE launched early versions of products into the market specifically to experiment (to gather customer feedback). Rather than relying solely on internal R&D, these firms experimented with preliminary offerings, learned from customer responses, adjusted their products and marketing strategies accordingly, and repeated the process for the next offering. In this way, experimental variation was not only intentional but also responsive—shaped and refined by real-time market input. Unlike traditional R&D efforts, this approach embeds variation directly into early commercialization, making customer feedback central to experimentation.

A similar dynamic is observed in teams using the lean startup method. Leatherbee and Katila (2020) for example show that lean startup teams often generate new hypotheses to test—not just as a starting point, but in response to feedback received during the previous experimentation cycle. Intriguingly, feedback can prompt entirely new avenues of variation (i.e., new problems to test), rather than simply validate or invalidate existing assumptions. Zuzul and Tripsas (2020) further reinforce this view, finding that external audience reactions can redirect the variation that founders pursue, making market feedback a critical and ongoing driver of experimental diversity.

Availability of technological and digital infrastructure can also affect experimentation. Advances in infrastructure can lower costs of experimentation, push teams to experiment, as well as reduce uncertainty through the use of predictive models (Brown & Eisenhardt, 1997; Raneri et al., 2023). The introduction of new information technology systems can also lead to experimentation, which in turn generate changes in the organization, such as the introduction of new work practices (Aldrich, 1999; Orlikowski, 1996). Together with technologically-aided experiments, experiments conducted via physical artifacts are also a relevant source of experimental learning.

A frequently studied external antecedent is uncertainty, which pushes teams toward experiments to reduce it (Berends et al., 2021; Murray & Tripsas, 2004). Market uncertainty is a particularly frequent stimulus that prompts the organization to test for new ideas. By experimenting, managers can explore the future by neither relying on a single plan nor being entirely reactive to external conditions (Nicholls-Nixon, Cooper, & Woo, 2000).

The regulative and normative mechanisms under which experimentation is conducted also emerged from our review as a relevant factor influencing experimentation. Within the French ecosystem, formal rules, incentives, values, and norms to which its members were exposed were shown to influence the adoption of a culture of experimentation via the lean startup method (Becker & Endenich, 2023). Peprah et al. (2022), in describing the case of Jumia in Africa, find that variation similarly emerges as a response to institutional voids—often ad-hoc and improvisational rather than intentional.

Within the field where organizations operate, competitive dynamics as well as other stakeholders' pressures can encourage or discourage experimentation as well. Studies find that experimentation can be mimetic. Firms scan the environment for new market trends and may adopt copycat strategies, which are later subject to experiments, to embrace new opportunities rather than search for novelty within organizational boundaries (Sanasi et al., 2022). Imitating established practices and business models has also been found to work effectively in developing economies (Peprah et al., 2022). Competitive pressure can increase the perceived usefulness of rapid experimentation and the likelihood of top management support since experiments are likely to be viewed as contributing to product improvements, competitive advantage, and performance (Zou et al., 2025). Beyond direct competitors, other stakeholders such as customers might influence organizational experimentation as well (Cui & Wu, 2017).

Finally, constraints at the field or industry level can significantly limit the degree of variation in experimentation. As Yaqub (2018) illustrates, vaccine development—particularly in areas like HIV research—is constrained by both the absence of viable testing models and stringent ethical considerations. These limitations severely restrict the scope and type of experiments that researchers can conduct. Similarly, in industries such as automotive safety, high-variation experimentation is only feasible through advanced simulation technologies, given the prohibitive cost and ethical implications of physical crash testing. In biomedicine, biological complexity and regulatory oversight further constrain the range of experimental variation. Experiments on organizational strategy or organizational structures are also often challenging and costly to implement. Across these domains, the nature of the "product", ethical boundaries, and technical feasibility combine to shape—and often restrict—how variation can be pursued in practice.

Experimentation programs. While the concept of experimentation programs is less frequently addressed in the problem-solving literature, a few notable studies provide insight into how variation can be deliberately structured to establish a “program” around experimentation. Harvey et al. (2023), for instance, argue that variation emerges from a team’s ability to switch fluidly among different learning modes: experimental (trial-and-error), contextual (recognizing patterns in the environment), and vicarious (learning from others’ experiences). Teams that skillfully balance these modes tend to introduce a broader range of approaches and variation into their experimentation efforts.

Effectively managing the volume of experiments is another critical element in programmatic experimentation. Organizations may face an almost unlimited number of potential experiments, but Gillier and Lenfle (2019) caution that too many concurrent experiments can overwhelm an organization’s capacity to learn or act. To address this challenge, Thomke and Bell (2001) propose a useful heuristic: the number of experiments an organization should run is proportional to the square root of their cost—balancing ambition with feasibility. Cost of experimentation is also related to other concerns. While online experiments are low-cost, if they result in negative customer experience, the costs can sky-rocket. In addition, in some industries such as automotive, experimentation requires costly and partially irreversible strategic commitments (Pillai et al., 2020).

Experimentation programs are also shaped by how experimentation is sequenced and paced. Brown and Eisenhardt (1997) describe successful firms as engaging in “semi-structured” experimentation, combining both simultaneous and sequential variation (see also Chen & Katila, 2008). These firms generate multiple design alternatives in parallel while maintaining flexibility to pivot as they learn. Importantly, they pace their experimentation efforts rather than rely on irregular bursts, which helps maintain momentum and adaptability in fast-moving environments.

Timing is another key consideration. Thomke and Fujimoto (2000) emphasize that experimentation programs are most effective when conducted early—before major resource commitments are made. Early experimentation helps innovators identify what works and what does not while the cost of change is still low, increasing the odds of success as ideas are scaled and implemented.

Selection

When experiments are run, what selection criteria are used to pick a winner? An intriguing finding in the problem-solving stream is that what gets selected is often shaped by individual preferences, identity, and interpretation, not performance metrics. In other words, identity is a strong determinant of what gets selected. For instance, in young firms, selection is often shaped by founder identity, emotional processing and internal sensemaking (Allen et al., 2024) and in bigger organizations by human preferences, voting and elimination rounds (Terwiesch & Ulrich, 2009). In the air taxi industry, for example, founders were more likely to continue with and support experiments that aligned with their identity frames, even when other ideas were more market-viable (Zuzul & Tripsas, 2020). Grimes (2018) similarly found that founders often discarded promising options when they conflicted with their self-concept, suggesting that selection criteria are filtered through identity coherence. Ventures that engage in identity and strategy work are also more likely to pivot toward paths aligned with their long-term goals rather than just interpret experiments on their own merits (Allen et al., 2024).

Leadership (especially the CEO) and other powerful internal actors also have a strong influence on how experiments are interpreted and what gets selected. Powerful internal actors often have a role in determining which variants survive; at Alessi, for example, aesthetic or functional logics were selectively adopted based on internal champions' preferences and evolving identity narratives (Dalpiaz et al., 2016). Other studies also widely document that perceptions of leader support in large firms shape which experiments the teams view as viable (Lee et al., 2004). This of course becomes problematic when leadership signals are inconsistent, and teams struggle to determine which paths are endorsed—undermining selection clarity. In the lean startup context, mentors improve selection quality by helping entrepreneurs interpret feedback and prioritize effectively (Leatherbee & Katila, 2020; Na et al., 2022; Motley et al., 2025). Overall, many studies in the problem-solving stream confirm that selection is not purely outcome-performance-driven.

Notably, the ability to select well appears to be a skill that can be both developed over time, and that can degrade over time, if unused. Using the V-S-R framework, Katila et al. (2017) find that medical doctors placed into CEO roles become less likely to select riskier, high-variance innovation paths, due to their atrophied skill to see the potential of high-variance alternatives. Ability to select viable alternatives from experiments, the authors found, is a practiced capability that needs to be continuously nurtured and supported—not something easily reactivated when needed. It is also interesting that the same individual may be highly skilled at planning for and imagining relevant variation for experiments, but terribly narrow-minded and biased when it comes to evaluating (and selecting) the winners of the experiments (Berg, 2016).

Cognitive diversity of the team (that supported the identification of variance in the first place) is surprisingly challenging for selection, and adds complexity. Schmickl and Kieser (2008) find that in interdisciplinary R&D teams, too much diversity can create cognitive distance that impedes convergence. Selection, in these cases, requires negotiation and shared understanding—not just evaluation. Over time, teams develop better interpretive skills and become more disciplined in how they evaluate alternatives, but short-term selection in diverse teams is challenging (Zhuge et al., 2023; Motley et al., 2025). Browder et al. (2022) provide a similar illustration from a cross-organizational (team) context. They show how selection in interorganizational settings is shaped by relational dynamics and interpretive alignment. Through structured dialogue, however, cross-organizational actors can collaboratively evaluate potential innovations—embedding multiple viewpoints into the selection process.

At the organizational level, selection is similarly portrayed as socially situated and interpretive. Success metrics are often ill-defined upfront; instead, selection occurs through a mix of rational analysis, organizational politics, and interpretive sensemaking. Selection is also influenced by broader narratives and identity goals of the organization. Goal alignment also functions as a strong filter for selection, especially in mission-driven contexts. Lynn et al. (1996) find that firms choose which experiments to continue based not only on market feedback, but also on internal judgments of strategic fit and future potential. In the Manhattan Project, Gillier and Lenfle (2019) document how technically promising designs were rejected if they didn't support the overarching goal of building the bomb—underscoring that strategic coherence often outweighs technical merit. Selection here reflects both bottom-up initiative and top-down integration with organization-level goals.

Despite the interpretive process described in many studies, selection is also described (in some work) to depend on performance. In more structured environments, selection benefits from simulation and prototyping tools. Thomke (1998) finds that firms equipped with such tools are better able to isolate causal mechanisms and select higher-performing designs. Effective selection, in this sense, depends on having the right infrastructure: rapid feedback, reliable testing, and supportive tools. Experience also plays a crucial role. Bruneel et al. (2010) find that firms with more international and technological experience integrate that experience into selection, and are argued to make informed decisions when selecting business model elements. Finally, selection capabilities also evolve at the firm level. Khanna et al. (2016) show that experienced firms refine their advancement criteria over time based on cumulative failures, gradually developing more sophisticated heuristics.

Selection, particularly for early-stage firms, is also strongly influenced by external validation, such that external forces sharpen selection heuristics over time. Becker and Endenich (2023) highlight the central role of user traction, revenue metrics, and mentor or investor feedback in shaping which hypotheses are pursued. In more mature contexts, external signals—trade shows, press coverage, and ecosystem endorsement—often determine what gets scaled (e.g., Dalpiaz et al., 2016).

Prototypes are a frequently used artifact for experimentation, which can be used for different purposes such as gathering knowledge on a predefined element or measure or seeking confirmation on selection from external audiences (Paust et al., 2024). Prototyping can also be performed in a collaborative fashion with internal and external stakeholders and has been shown to lead to effective prototype revisions (Bogers & Horst, 2014) and possible selection of ideal product specifications (Terwiesch & Loch, 2004).

Experimentation programs. At the programmatic level, firms improve selection through complementary approaches to experimentation. Posen and Levinthal (2012) argue that combining vicarious learning with local experimentation improves selection by balancing exploration and efficiency. In robotics R&D, Katila and Ahuja (2002) similarly find that exploration and exploitation operate in distinct but synergistic zones of the search space, enabling better-informed selection.

Finally, structural experimentation—such as running multiple business models in parallel—can enhance selection. Andries et al. (2013) find that startups engaging in simultaneous experiments are better equipped to evaluate which configurations are most viable, underscoring how structural variation supports more effective selection.

Retention

In the problem-solving stream, retention refers to how results of experimentation are preserved and embedded within the organization. Ideally, successful results and elements from experiments are institutionalized into long-term development strategies, technical standards, collaboration norms, and everyday practices—ensuring that knowledge persists beyond the immediate experiment (Lynn et al., 1996; Harvey et al., 2023). Reflecting this optimistic view, Alblas and Notten (2021) describe how debugging knowledge can be retained by incorporating checklists, problem-solving heuristics, and "debug playbooks" into engineering routines, allowing future teams to benefit from past insights. Indeed, some studies present a positive view of retention, emphasizing how lessons learned are codified into reusable design principles, simplified prototypes, and organizational routines (e.g., Gillier & Lenfle, 2019).

Experimentation inside an organization can become a program of experiments that unfolds over time rather than as a series of one-shot tests. Such nexus is described by Paiola et al. (2022) who study the evolution of IoT-enabled business model innovation which takes place through incremental, trial-and-error adjustments. Whereas initially experimentation activities are carried within the same business units, at a more advanced stage, the team in charge of experimentation is bought, in one instance, out by the parent company and incorporated in an independent business unit.

However, retention is not guaranteed. Fast experimentation cycles and short-term pressures can hinder or delay reflection, making it less likely that valuable insights from experiments are captured or reused over time (Alblas & Notten, 2021). In such cases, organizations may rush from experiment to experiment without pausing to consolidate what was learned—especially in rapid product development environments. This not only weakens retention but may discourage experimentation altogether. In fact, the emphasis on the speed of experimentation and rapid validity testing (Pena & Haufler, 2021; Zou et al., 2025) may have resulted in more limited research on whether and how the learning acquired via experimentation is retained over time. For example, it remains unclear whether participants in structured programs (e.g., 8-week training on experimentation) integrate these practices into ongoing routines or treat them as one-off experiences.

Relatedly, retention can backfire. Andries et al. (2013) note that prevailing cognitive commitments and organizational inertia may lead firms to select and subsequently retain the familiar but suboptimal elements from experiments, even when experiments in their entirety reveal better alternatives. Even when experiments disconfirm an idea, identity-consistent elements may survive and become embedded in the venture's trajectory, trickling down and contributing to organizational rigidity in the long run. In this sense, retaining the wrong lessons can be worse than retaining none at all.

A few studies also point out that organizational structures and designs can serve as powerful vehicles for retention (Reineke et al., 2025). Experimental insights can be embedded and retained through roles and structures, institutionalized into the firm's structure, reinforcing its identity and strategy. At Alessi the new role of design assistant combined responsibility for the efficient completion of projects with preserving the integrity of designers' ideas; design by meta-project replaced traditional market research with social science workshops; and books and texts institutionalized and preserved the vocabulary of organizing (Dalpiaz et al., 2016). Another relevant organizational design element is leadership support and how leadership perceives success. When experimentation is viewed as successful, organizations are more likely to retain the associated behaviors, even after a crisis (Zhuge et al., 2023).

In emerging markets, retention can reflect aggregation based on adaptation to local contexts. Peprah et al. (2021) illustrate how Jumia retained context-specific solutions by institutionalizing them after repeated rounds of experimentation. These retained practices became part of the firm's operational DNA, ensuring relevance and responsiveness in complex environments.

Again, retention can also be influenced by the sequencing of experiments. Firms conducting simultaneous experiments can gain more comparative insight, enabling more deliberate and informed retention of superior options (Chen & Katila, 2008).

Summary. Experimentation in the P-S stream is iterative and pragmatic, aimed at relevance and feasibility rather than isolating a single causal ingredient. Variation can be blind and emergent (sparked by ad-hoc ideation, anomalies, or deviance, and supported by decentralization), or more intentionally structured through tools like prototyping, and supported by strategic leadership. Individual, team and organizational and ecosystem factors influence what variation gets generated. Selection is often interpretive and socially situated: preferences, power, and sensemaking shape "winners" as much as performance metrics, though experience can improve selection quality. Retention ranges from codifying lessons into routines and roles, to weak and even harmful retention, where speed pressures and inertia cause organizations to keep familiar but wrong lessons.

Causal inference stream

C-I studies frame experimentation as tightly scoped tests designed to estimate unbiased causal effects of a well-specified intervention. Although online controlled experiments are a rapidly rising category today, C-I studies are prevalent in many different sectors besides online platforms. We again organize the literature using the classic variation–selection–retention (VSR) framework.

Variation

What is the genesis of variation, i.e., how are problems framed in the C-I stream? Again, variation arises through multiple mechanisms.

Blind versus intentional variation. In contrast to the more ad hoc, "blind" forms of variation underlying the problem-solving experiments, variation in the causal inference stream is highly systematic and intentionally designed. In fact, most causal inference research is skeptical that blind variation can bring any value. Blind variation is typically treated as an error (in coding, analysis, interference, violating assumptions, etc.) to be eliminated (Kohavi et al., 2020) rather than an anomaly (Thomke and Loveman, 2022) that could lead to discovery of new customer demand or delight. It even has a label, Twyman's law, widely embraced by online controlled experiments research to indicate that anomalies are typically not meaningful. According to this law, "Any figure that looks interesting or different is usually wrong" (Kohavi et al., 2020: 39).

In a typical causal inference (controlled) experiment, users are randomly split between the variants (e.g., two different ad layouts) in a consistent manner over time (e.g., a user receives the same experience in multiple visits). User behavior such as interactions with the site are tracked and key metrics computed. The core tenet of causal inference experiments is to reliably identify causality with high precision, e.g., the features that cause changes in user behavior (Ejdemyr et al., 2024). At the core of causal inference experiment is the variants that are being tested. These experiments can be run if treated (variant) and controls are separate groups and do not influence one another; enough experimental units are present; and changes can be made relatively easily. All of these conditions limit the variants that can be tested in practice. Dominant platforms like Google and Amazon exemplify this approach, often varying product features deliberately to test and refine their understanding of customer preferences as well as to improve their firm-specific strategic and competitive considerations. For example, Kim and Luca (2019) and in a rebuttal, Millet, Lewis and Stoddart (2022) examine different types of reviews that are displayed to users of a search engine as a response to queries.

In the causal inference stream, variation is approached with structure and precision. Kohavi et al. (2013) document Microsoft's A/B testing that often involves small, controlled changes—such as tweaking messaging or interface design—to isolate causal effects. Cooprider and Nassiri (2023) describe Amazon's price experimentation via switchbacks which repeatedly cycle a product between a treatment and a control condition over time. Much of the work we reviewed emphasizes the importance of minimizing "confounding variation" by isolating testable components and structuring experiments to cleanly identify what works (e.g., Holtz et al., 2020). Further, variation in the causal inference stream extends beyond user interface or content changes. Platform firms vary algorithms, operational procedures, and pricing strategies—expanding the domain of what can be experimentally manipulated.

Recent causal inference studies emphasize that variation is not only intentionally designed by teams of engineers, but also increasingly data-driven. AI (machine learning) can be used to guide where variation should occur, identifying promising areas for intervention based on algorithmic predictions. This "algorithmic variation" allows firms to direct experimentation efforts toward areas with the highest potential payoff (see also Rathje, Katila, and Reineke, 2024).

A/B testing infrastructure, for example, has routinized the process of generating variation: firms can systematically test alternative product features, user interfaces, pricing structures, and algorithms in parallel (Koning et al., 2022). As a result, variation becomes not just intentional but scalable—transforming experimentation into a reproducible organizational capability. Indeed, findings show that startups conducting more A/B tests generate more performance-relevant variation, and, are more likely to identify superior alternatives (Koning et al., 2022). Research particularly encourages high-throughput variation which encourages a culture of many concurrent experiments to produce a constant inflow of variation. As an example, Netflix introduces variation at massive scale through its experimentation platform, running thousands of concurrent experiments (Ejdemyr et al., 2024).

Individual factors underlying variation. Despite its scientific precision, many C-I studies document individual-level drivers of variation. First, leadership buy-in is an often-mentioned driver (or lack of it, an impediment). A particular individual's opinion, labeled as HiPPO (highest paid person's opinion), often has an outsized influence on experimentation in platform organizations. Reflecting their experience at Microsoft, at Google and at LinkedIn, Kohavi and colleagues (2020, p. 60) note that "Leadership buy-in is critical for establishing a strong culture around experimentation and embedding A/B testing as an integral part of the product development process…organizations and cultures go through stages…the first stage which precedes any experimentation, is hubris, where measurement and experimentation are not necessary because of confidence in the HiPPO (Highest Paid Person's Opinion)…Next, an organization starts measuring key metrics and controlling for unexplained differences…Last stage…where causes are understood, and models actually work." It is in this last stage, the authors write, that leadership buy-in for experimentation becomes critical, including (1) agreeing on high-level goal metrics and guardrail metrics; (2) setting goals in terms of improvements to metrics instead of goals to ship features X and Y (including not shipping a feature unless it improves key metrics); (3) enforcing standards on interpreting results and giving transparency to how results affect decision-making; and (4) ensuring a portfolio of high-risk/high-reward projects as well as incremental ones (Kohavi et al., 2020). All of these factors are significant inputs of leadership to the process. Phillips (2025) notes the same issues at Uber and at Amazon. Leadership's individual beliefs of what should be tested and what not, do trump experimentation programs, even in big firms, so careful alignment of experimentation programs with firm and leadership strategy is essential.

Studies we reviewed also describe how experiments may sometimes threaten an individual manager's expertise and may be related to "perceived loss of power" and so may be resisted (e.g., Kohavi et al., 2009). At Amazon, the data science team created a prototype of personalized recommendations based on items in the user's shopping cart. Despite the promise, a marketing SVP was "dead set against it" claiming it would distract people checking out, and the data team was forbidden to work on it further. But a controlled experiment showed that the feature won by such a wide margin that not having it was costing Amazon a "huge chunk of change" and was eventually adopted (Kohavi et al., 2020).

Another frequently mentioned antecedent is employee and platform-complementor buy-in (versus resistance). This is critical, as experiments that are concealed from complementors, for example, can end up hurting the platform in the long run as they can alienate both users and complementors (Rahman, 2024; Marres & Stark, 2020) or cause user outages (Feng & Zhao, 2017).

Organizational factors underlying variation. An important organization-level determinant of variation is experiment's potential to generate business impact. Whereas curiosity often prompted problem-solving experiments, in the causal inference stream, experiments are most often driven by their ability to boost financial performance. These criteria are particularly challenging as most experiments fail: at Slack, only 30% of monetization experiments showed positive results and other organizations put the success percentage close to 10% or even 1% of all experiments that are run (Kohavi et al., 2020). Therefore, it is difficult to justify adding variation, and experiments, as most of them do not generate positive performance impact.

Cost-benefit calculus of experimentation also matters. Not all experiments are worth running—particularly those with low expected impact or high implementation cost. The literature also explores what determines whether organizations engage in "negative experiments" (i.e., those that fail to outperform the control and thus end up hurting customer experience during the experiment). Despite the cost, they may emerge as important sources of learning, reinforcing the disciplined approach that underlies variation in the causal inference tradition, yet they may not be run given the performance benchmarks.

Once an experiment is approved, a key question is which organizational unit determines the exact variation to be tested versus who runs the experiment. At Uber's 'Experimentation Labs', governance rules specified that the lab itself selected which experiments to run, drawing on ideas submitted by middle and product managers. Importantly, the variation to be tested was expected to originate from a group external to the lab, reinforcing a separation between idea generation and execution (Phillips, 2025). Organizationally, in causal-inference experimentation, the ideas for the incremental variations should ideally originate outside of the data science teams running the experiments. Sourcing variation (ideas) from product managers or other functions helps ensure that the experiments are not overly driven by data patterns alone, and helps guard against biases such as "peeking" at results midstream. The lab's task is then to prioritize experiments with potential for high business impact and long-term value. The lab also explicitly deprioritizes exploratory or 'what-if' experiments unless a clear business case could be made.[5] This process is typical of many platform companies running online experiments.

Organizations' ability to master several different types of data science methods besides experiments is also linked to origins of variation. Allen and McDonald (2025) argue and find that methodological pluralism (quantitative, qualitative, hybrid) facilitates richer variation in problem framing. Drawing on a dataset of new-product innovations linked to employee resume job descriptions (to track methods capabilities used in the firms), the authors find that relying primarily on either type of analysis has little impact on overall new-product sales but together has a positive effect as each compensates for the weaknesses of the other. Amazon offers another example. When Prime Video, originally introduced as Amazon Unbox, did not meet

[5] Uber is an interesting example of causal-inference-driven experimentation. See for instance https://www.uber.com/blog/supercharging-a-b-testing-at-uber . Wu, Lou, Hitt (2019) also examine the role of decentralized structures to make the most of algorithmic experimentation: They find that the ability of analytics to link knowledge across diverse domains and to integrate external knowledge into the firm is its main benefit (i.e., create new combinations and reuse of existing technologies).

performance expectations, Amazon conducted multiple *types* of experiments — including A/B testing across different markets, such as bundling Prime Video in one region versus offering it separately in another, alongside segment-level cohort analyses — and through the mix of methods discovered that users were more likely to subscribe to Prime when video was included as part of the Prime subscription package. In parallel, Bruneel, Yli-Renko and Clarysse (2010) show that variation for experimentation is often amplified by external relationships (e.g., universities or customers), bringing in novel knowledge and variation.

External factors underlying variation. Few studies discuss external determinants of causal-inference experiments. Focusing on the regulatory environment as a predictor of an organization's use of A/B testing, Zou et al. (2025) find that A/B testing is perceived as easier to adopt when the regulatory environment provides better guarantees and support.

While much focus of current experimentation literature is on A/B tests, external factors may also influence variation in a sense that in many settings finding a comparable control group is a challenge (e.g. in two-sided markets traditional A/B tests do not produce reliable results because you cannot experiment on both sides). In such situations (airline flight pricing is another example), "switchbacks" is a more common way of testing, akin to sequential reorientations (McDonald & Gao, 2019; Bojinov et al., 2020, 2023).

Selection

In the causal inference stream, selection is framed as a systematic, data-driven process grounded in rigorous quantitative analysis and driven by p-values, metrics, and quantitative data. Randomized controlled trials are the gold standard for establishing causality, and the randomization process is central in selection, to ensure that the populations assigned to the different variants are similar statistically, allowing causal effects and winning variants to be determined with high confidence.

Much recent research in the causal inference stream underscores how organizations operationalize selection at scale. For example, Kohavi et al. (2020) outline several methodological principles in large-scale web experimentation, including selecting variants based on statistically valid performance metrics (e.g., click-through or engagement rates), avoiding “peeking” bias to ensure validity, and automating selection through real-time metrics on experimentation platforms. Similarly, Ejdemyr et al. (2024) show how platforms such as Netflix uses precise thresholds to determine selection. Treatments are chosen only if they produce a measurable lift in key outcomes such as engagement or retention (Rahman, 2024). This quantified decision-making, codified into platform rules, sets a high bar: only a small fraction of experiments meets the threshold for implementation. Koning et al. (2022) similarly reinforce that A/B testing enables formal, scalable selection. Firms that run high volumes of experiments and use conversion-based metrics are more likely to select beneficial variants and improve performance over time. In sum, rather than the mostly subjective, political or ad hoc selection seen in the problem-solving stream, causal inference studies portray selection as a replicable, data-validated process.

Many data-driven organizations have established rules for selection. At Uber ‘Experimentation Labs’, for example, the rules include (1) no experiments run outside the data team – to ensure that no one running the experiment had a stake to influence the interpretation of the results; (2) no metrics hacking – i.e., headline metric is specified upfront, and no changes to metrics or time frame are allowed during the experiment (Johari et al., 2015); and (3) no slicing and dicing (Kohavi & Thomke, 2017).

Selection becomes more robust when organizations use multiple modes of experimentation. Cross-method triangulation—actively comparing results across different models—enables more reflective and rigorous selection, moving beyond reliance on any single data stream or model output (Allen & McDonald, 2025).

Retention

In the causal inference stream, retention refers to how organizations document, reuse and build further experiments on the insights generated through past causal-inference experimentation. C-I research emphasizes that identifying the mechanisms underlying the results of the experiment enables generalization and longer-term impact. Kohavi et al. (2013) highlight the importance of recording not only the outcomes of experiments but also the decision rationale behind them. This documentation enables future experiments to build on prior learning and helps prevent redundant testing of previously rejected ideas. Organizations attempt to institutionalize experimentation knowledge through internal algorithms, performance dashboards, and curated playbooks, which serve both as repositories for lessons learned and as training tools for future experimenters (Ejdemyr et al., 2024). These efforts represent early steps toward building a systematic memory for experimentation.

Despite the ideal, studies suggest that real-world implementation of experimentation "programs" that retain and build knowledge repositories still remains limited. Many platform organizations note that retention of what was experimented relies more on "tribal memory" than any systematic retention (Phillips, 2025). Fast-moving business priorities, shifting success metrics, and a prevalence of one-off tests pose significant challenges to building cumulative experimentation capabilities. As is well-documented, most causal inference experiments—upwards of 99%—are not ultimately implemented. However, when experiments do identify successful interventions (roughly 0.5% of the time), these changes are often scaled and embedded directly into product features.

Finally, an important consideration is the scope and boundary conditions of experimentation. Wu, Hitt and Lu (2019) find that data analytics capabilities (measured by resume data) are more likely to be present and are more valuable in firms that are oriented around incremental improvements than they are in firms that are focused on generating entirely new technologies, suggesting that retention of experimentation capabilities is more valuable in experiments that focus on incremental rather than radical improvements.

Summary. Experimentation in the C-I stream is tightly scoped and deliberately narrow. Variation is intentional, structured, and increasingly high-throughput and data and AI-guided, with anomalies treated as errors to eliminate rather than sources of discovery. Individual and organizational factors with a strong cost-benefit focus shape what gets tested. Selection is portrayed as a rigorous and metrics-driven but can still be influenced by organizational politics and power. Retention emphasizes documenting results and rationales to enable cumulative learning, but in practice remains limited and often relies on tribal memory, with successful insights mainly supporting incremental rather than more radical change.

Outcomes of problem-solving and causal-inference experimentation in organizations

A final significant set of findings focused on the effects of experimentation and experimentation programs on other, broader organizational outcomes such as financial performance and long-term survival. Other consequences that were examined in the articles we reviewed include learning and knowledge acquired via experimentation; innovation in its different forms including new product development; growth; and, adaptation and change, such as pivoting the entire business model or parts of it. In reviewing and analyzing this evidence, we step back and ask the question: does experimentation matter?

Overall, research suggests that experimentation does matter. Consequences of experiments in organizations are on balance performance-enhancing: they offer the opportunity to test alternative business models (Khanagha et al., 2014) and refine existing ones (Burnell et al., 2022), improve survival and follow-on funding opportunities for startups (Motley et al., 2025), and create opportunities for organizational learning (Koporcic et al., 2024) and revenue (Lukkarinen & Katila, 2026). Causal inference experiments are also found to be performance-enhancing. Koning et al. (2022), for example, tracked 35,000 global startups, and found that

A/B testing lead to a 10% increase in visits in the first few months after adoption and after a year of experimentation, to 30% to 100% increases.

In both the problem-solving and causal inference streams, research cautions against focusing solely on first-order effects while neglecting second-order ones. For instance, in the problem-solving stream, studies warn that short-term outcomes can be misleading if longer-term dynamics are not considered. Thomke and Loveman (2022), for example, describe Caesar Palace’s introduction of a virtual queue system for its Bacchanal Buffet in Las Vegas. Initially, the experiment caused some customers to leave because they were no longer tethered to the restaurant's physical location while waiting. Yet, the second-order effect turned out to be highly beneficial: by notifying customers via text when their table was ready, the system allowed them to spend time elsewhere—often buying drinks or gambling—leading to additional revenue that more than offset any initial loss in diners, emphasizing the need to compare both short- and long-term outcomes. Similarly, in the causal inference stream, research cautions over-emphasizing first-order effects. For instance, platform interface changes that slow down the search process may lead to an increase in search queries in the short term (a positive first-order effect), but ultimately drive users away from the platform over time (a negative second-order effect).

A natural consequence of experimentation is related to the experience and knowledge that those involved in experimentation can accumulate. Through what Paust and colleagues (2024) describe as flexible experimentation, entrepreneurs formulate some minimum criteria or success measures, which they use repeatedly in their experiments to gather knowledge without following the more rigid structure suggested in studies embracing the theory-based logic (e.g., Camuffo et al., 2020). Other studies document that by using several approaches to experimentation, such as A/B testing and minimum viable products, managers gather and assimilate new knowledge and build their capabilities in experimentation.

Another organizational consequence of experimentation is the transfer of knowledge, which can be facilitated by translational mechanisms such as sketches, visualizations, and prototypes (Corbo et al., 2020). More innovative projects, for instance, have been found to produce more intensive knowledge transfer between the specialists of an organization due to higher number of iterations needed to identify errors via prototyping (Schmickl & Kieser, 2008). A commonly-adopted approach in experimentation within younger organizations is the so-called "fail fast, learn fast" mindset that encourages rapid testing and iteration to quickly identify viable solutions. Although more prominent in the startup culture, and particularly entrenched in ecosystems such as the Silicon Valley, established businesses have been shown to learn from failures in experimentation as well (Khanna et al., 2016) following a process that involves failure recognition, interactive sensemaking, and organizational adaptation (Koporcic et al., 2024). Studies focusing on problem-solving find that a problem-solving experimental approach positively affects those who adopt it, for instance, by increasing their self-efficacy and confidence (Mansoori et al., 2019).

Studies linking experimentation to performance generally rely on the prediction that organizational experimentation is performance-enhancing (Katila & Ahuja, 2002; Bingham & Davis, 2012; Camuffo et al., 2020). This is because through experimentation, organizations learn, and such learning should drive improvements in performance. Some empirical research provides confirmation. For example, Thomke (1998) tests the relationship between design performance and experimentation strategy and finds that through iterative, well-managed experiments, design teams learn from feedback and enhance overall product performance. When performance is below aspirations, organizations experiment through nonlocal search (Cyert & March, 1963). App development organizations, for example, have been found to engage in a nonlocal search strategy when their first product's performance is low. Yet, this experimentation strategy is found to be less beneficial than a moderate search strategy (Angus, 2019). Studies within the causal inference stream confirm that the adoption of a scientific method increases performance measured as revenue growth. Ideas where the entrepreneurs behave like scientists have a higher expected value given the more stringent conditions for developing an entrepreneurial idea (Camuffo et al., 2020). Several authors (Thau et al., 2014; Allen & McDonald, 2020) advocate the idea of methodological pluralism – such as the adoption of both quantitative (e.g., A/B testing) and qualitative (e.g., focus group) analysis in organizations – and suggest that combining several methods balances their strengths and weaknesses, ultimately boosting performance through their complementary effects. In contrast, McDonald and Gao (2019) study robo-investing ventures and their sequence of reorientations (sequential experimentation strategies), and find that not all experiments are performance-enhancing. Reorientation without penalty may depend on how the ventures anticipate, justify, and stage changes to various audiences. Knowledge repositories and routines as well as other tangible assets have been shown to improve experimentation effectiveness (Konietzko et al., 2020) and performance (Harms et al., 2020). Finally, the complete lack of experiments can be detrimental, as demonstrated by the Challenger disaster (Hauptman & Iwaki, 1991). Tesla and Samsung are examples of companies that have tried to discontinue experimentation for particular, mature products, with negative consequences.

Another important consequence of experimentation is pivoting, i.e., the intentional strategic reorientations taken by managers with the aim to increase convergence between the organization and external pressures (Sarta, Durand, & Vergne, 2021). The pivoting concept is well-rooted in theories of change and adaptation (e.g., Aldrich, 1979; Helfat et al., 2009). As organizations experiment, they gain knowledge from successful and less successful tests, which then inform the next experiments. Within this process, aspects that do not fit with the firm's strategy become a target of correction or elimination (Motley et al., 2025). Pivoting is particularly embraced when new information conflicted with or expanded the beliefs (Kirtley

& O'Mahony, 2023) and when organizational identity was clarified and encoded in the roles of organizational members through role crystallization (Snihur & Clarysse, 2022). Andries et al. (2013) documented the evolution of six organizations in their early stages and found that those that pivoted (i.e., did not commit to a specific business model too early but instead developed diverging search paths through business model experiments) were more likely to achieve long-term survival. Camuffo and colleagues (2020) similarly show that experimentation improves the precision of the teams, leading to better performance and a higher likelihood of further experiments such as pivoting to a different idea.

DISCUSSION AND CONCLUSION

In this integrative review, we took stock of research on experiments by organizations. Our analysis identified two streams: problem–solving–based experimentation, which emphasized iterative testing and learning from failure, and causal–inference–based experimentation, which emphasized experiments as tightly-scoped interventions to estimate causal effects. Within each stream, different experimental types (e.g., A/B testing, prototyping, switchbacks) emerged, which we formalize in Table 2. Four major insights emerge from our review.

[INSERT TABLE 2 ABOUT HERE]

Major Insight #1: Complementarities and differences: Problem-solving vs. causal-inference experimentation

The problem-solving and causal inference streams offer two distinct logics for designing, interpreting, and institutionalizing experiments in organizations. While both aim to inform and support decision-making, they diverge in *goals* and underlying assumptions (see Table 3). Problem-solving experimentation centers on generating novel solutions to messy, ill-structured challenges whereas causal inference experimentation focuses on identifying the outcome of specific interventions using well-defined metrics. In other words, the former seeks open-ended exploration and discovery and is well-equipped to support path-breaking innovation and change while the latter aims for precision and optimization, sharpening and scaling the current path.

A second key distinction between the two streams is the source and structure of *variation*. Under the problem-solving logic, variation often emerges organically—from improvisation, failures, team dynamics, or heuristics shaped by local knowledge and practice. Experiments are frequently initiated without pre-specified variables or clear hypotheses, relying instead on emergent learning. By contrast, causal inference experiments emphasize rigor and structure: variation is intentionally introduced in a controlled manner, such as through A/B testing, to isolate cause-and-effect relationships systematically. These experiments are pre-defined and tightly scoped to test hypotheses about variations such as in feature changes, pricing adjustments, and user behavior. Another key difference is the role of failure. In problem-solving, failures are generative. They serve as triggers for new rounds of search, reflection, and learning. In causal inference, in contrast, failure is embraced only to the extent that it supports statistical inference. The emphasis is on minimizing noise and confounds: failed variants are discarded unless they show a significant lift in performance metrics. Overall, the genesis of variation supports the distinct goals of the two approaches: exploration versus refinement.

Another central difference across the streams is *selection* and *retention*. Under problem-solving, selection is interpretive, and what gets retained is often filtered through social and cognitive lenses: founder identity, organizational culture, symbolic meaning, or practical routines influence what insights endure. Retention is often embodied in roles, product forms, or long-term identity narratives. By contrast, under the causal inference logic, selection and retention, at least in theory, are de-personalized and data-driven. Although power and politics play a role, the intent is that winning variants are embedded into codebases, dashboards, or platform algorithms with minimum human evaluation or involvement. Knowledge is thus institutionalized through experimentation infrastructure, which prioritizes reproducibility and scalability over contextual or cultural richness. Again, the focus on precision over occasional deviation is consistent with the causal inference stream's logic relative to the problem-solving stream.

Finally, the two streams differ in their experimental units, processes and infrastructure, again reflecting the goals of each approach. Problem-solving experiments are more often team- and project-based. They often lack formal codification, relying instead on informal coordination and situated judgment. In contrast, causal inference experiments are often embedded at the product, user, or algorithm level—enabled by highly formalized systems such as experimentation platforms. These differences reflect divergent philosophies: one rooted in creative search and ambiguity, the other in precision engineering and statistical control. Together, they illustrate the diverse ways organizations can learn through experimentation, and the tradeoffs between exploration vs optimization embodied in the two experimentation approaches.

[INSERT TABLE 3 ABOUT HERE]

Major Insight #2: Validity of inference

Our review also yields an opportunity to compare how quality indicators of experiments are prioritized and the trade-offs made across the streams. In particular, we assess the validity of experiments (the approximate truth) drawn from experiments in each stream using four types of validity (internal, external, construct, statistical conclusion) and highlight the threats most salient and most likely in each stream.

Internal validity refers to the credibility of an experiment's causal claim within its own setting—i.e., the extent to which observed differences in outcomes can be credibly attributed to the intervention rather than to other factors. Did the treatment cause the difference, not the confounds? In the causal inference stream, internal validity is at the core. Experimentation seeks clean counterfactuals via randomization and tries to account for salient threats like contamination. In the problem-solving stream, in contrast, internal validity's role is more peripheral: it matters insofar as the results are decision-useful for improving a product or a decision. However, studies in the problem-solving stream do not prioritize precise identification. Because focus is on iteration and practical problem-solving, many classical threats to internal validity (concurrent changes, selection into treatment, spillovers, etc.) can be present. Indeed, in contrast to tightly-scoped C-I experiments, the immediate threat is the holistic approach: P-S experiments often come bundled, making it unclear which ingredient mattered.

The next type of validity, external validity, is defined as the degree to which experiments' findings and the estimated causal effect generalize beyond the specific setting used in the experiment. While external validity is desirable in both streams, it is typically secondary to solving the immediate problem the experiment is designed to solve, particularly in the causal inference experiments. In the problem-solving stream, external validity matters more (e.g., scaling a locally useful, context-specific fix to other teams, products, or business units), but it is not an explicit focus. In the causal inference stream, while external validity is discussed (e.g., internally-valid results would not automatically travel across segments, channels, or time), in practice experiments almost always rely on selective settings and carefully designed but narrow experiments, limiting generalization.

Third, construct validity is the degree to which the manipulation (treatment) and the operationalizations truly capture the theoretical ideas they are claimed to represent. Is the treatment what we think it is, so we can explain *why* it worked? Again, construct validity is not a priority in either stream. In problem-solving, because the teams frequently iterate both the manipulation and the metric, treatment and construct boundaries often are blurred, making it easy to attribute an effect to the wrong construct. Nevertheless, construct validity is an important concern for implementation. In contrast, while C-I experiments target constructs more explicitly, they often misfire if measures are sterile and distant from constructs (e.g., click-through for "value" or "interest"; Lukkarinen & Katila, 2026). The disconnect is frequently present because most organizations run on short-term goals: understanding the mechanisms (and the construct) behind the results is often less important than the immediate result. If one-word change in the website description changes the revenue by $0.5M but the effect is likely to be temporary, understanding why it happens may take back seat, and rapid implementation front seat. While mechanisms still matter, in fast-moving environments, construct validity may not be the main concern for the organization, especially in the causal inference stream.

Fourth, statistical conclusion validity asks whether there is an effect (not noise), i.e., the extent to which the statistical analysis justifies the inference that an effect exists, given the data and procedures used. In plain terms: do the numbers support the conclusion, with acceptable error rates and precision? Causal inference experiments prioritize statistical conclusion validity. In the problem-solving experiments, in contrast, more threats to validity exist. Overall, each streams make trade-offs: experiments in the causal inference stream prioritize statistical conclusion and internal validity while experiments in the problem-solving stream place relatively more emphasis on construct and external validity.

Major Insight #3: An Integrative Conceptual Model

The problem-solving and causal-inference-based approaches to experimentation are also reflected in the design choices implemented at each stage of organizational experimentation (see Figure 3).

The causal inference research emphasizes *variation* as being business-, data-, or leadership-driven, while the problem-solving stream highlights less structured and spontaneous ideation rather than intentional planning. In parallel to more traditional ways of experimenting, such as prototyping, variation is also embraced more informally through organizational storytelling, which has also been found to help organizations generate and test assumptions.

The *selection* of the most promising variants in problem-solving experimentation occurs through a combination of rational analysis and interpretative sense-making. Hence, a more structured approach at this stage via prototyping and simulation leads to more effective selection. In causal-inference-based experimentation, where the selection is based on data and rigorous analysis, RCTs represent the gold standard followed by A/B tests and switchbacks which enable further testing between alternative options.

Retention and consolidation of successful results into standardized procedures and routines is an equally important but less explored aspect in the literature. Through retention, successful variations are kept and eventually become part of the organizational repertoire. This process of institutionalization of best practices of experimentation is desirable as it allows the organization to more easily reproduce experiments effectively and learn from them within and across organizational units.

[INSERT FIGURE 3 ABOUT HERE]

Figure 4 presents an integrative conceptual model of the drivers, processes, and consequences of experimentation derived from our review. In particular, it identifies (1) the common drivers of experimentation (individual/team, organizational, and external) that frame the variation to be tested; (2) the two key processes, problem-solving and causal-inference-based experimentation, that are implemented through the stages of selection and retention; and (3) the major organizational consequences related to innovation, performance, and pivoting.

While the studies in our review suggest that learning is a key element of experimentation, our model helps generate new insights on whether organizations *actually* learn, and, even more importantly, whether there is commitment to abide by the results of the experiment, a point often emphasized by problem-solving work (Thomke, 1998, 2020). One observation, for example, is that most learning from experiments is incremental: as firms gain experience, trial-and-error processes engender the refinement of ever-more efficient routines, de-emphasizing change and innovation (Thompson, 2010). Moreover, the literature focusing on pivoting has generally explored the possibility that organizations change one or more aspects of their business following what they observe and learn through experiments. However, these reorientations are often viewed as single moments in the time of the studied organizations, despite the fact that in practice they are often part of a broader experimentation program. The model we develop helps illustrate the possibility of *feedback loops* that result from repeated experiments. One possibility is that an experiment leading to unsatisfactory results pushes the organization to start new experiments, alternating between problem-solving and causal-inference-based experiments. Instead of viewing learning as the natural consequence of experimentation, organizations may undergo a process of international unlearning before starting new experiments. Another possibility is that, even in the face of successful experimentation, an organization decides to continue experimenting, establishing a culture of experimentation where curiosity and surprises are valued more than the costs inherent in experiments (Thomke, 2020).

[INSERT FIGURE 4 ABOUT HERE]

Major Insight #4: Does The Ideal Experimental Organization Exist?

A key question is whether organizations can develop over time into "master experimenters", eventually approaching an *ideal* of experimental organization. Drawing from our review, we propose that organizations and their experimental approaches co-evolve along two key fundamental dimensions: the level of experimental maturity and the scope of experimentation.

The level of experimental maturity distinguishes organizations based on their cumulative experience with experiments. Early experimenters can be both young organizations with little experimentation experience or established organizations embracing experimentation at a later stage in their life cycle. Mature experimenters, instead, have more consolidated experience and have embedded experimentation as part of their overall organizational processes.

The scope of experimentation is the degree to which experimentation is limited (typically R&D, product features, manufacturing, marketing) or extends more broadly across different functions of the organization (horizontally, vertically, or both). At present, few or no empirical studies address experiments with different organizational designs, people processes or corporate strategies, for example. The combination of maturity and scope gives rise to four types of experimenting organizations (see Figure 5).

We define organizations that have a lower level of maturity with and a narrow scope of experimentation as 'proactive' in the sense that they mainly use experimentation as a means to limit short-term uncertainty and to address challenges before they become critical. Proactive experimenters (QI in Figure 5) would typically embrace one-off experiments to address time-bound uncertainties.

In contrast, organizations that have reached a higher level of experimental maturity but still perform most if not all experiments locally (e.g., R&D departments) typically adopt a sequential approach. For instance, engineers aiming to fix a problem in a manufacturing process might first isolate the location of the cause, formulate hypotheses, and then test possible interventions to find whether or not to fix the problem (Bohn & Lapré, 2011). We define such organizations as 'reactive' (QIIa) since experimentation is generally response-based and driven by organizational challenges or external stimuli.

Early experimenters who perform experiments at a broader level within the organization transition from 'proactive' to 'structured' experimentation (QIIb). Given the increasing complexity stemming from expanding testing from a restricted group of individuals (e.g., a product team) to a broader number of organizational members (e.g., business units), organizations may slowly move from trial-and-error to more controlled and replicable testing embracing a causal inference logic.

Finally, organizations embracing experimentation extensively and with more experience correspond to stylized imagery of an "ideal experimenting organization" and are defined as "systematic" experimenters (QIII). Such organizations view experimentation as a core element driving decision-making. Employees, regardless of rank or unit, are encouraged to frame hypotheses, run trials, and draw lessons to inform and guide decision-making. In this context, experimentation is not confined to a single function, but a central operating system that shapes strategy, product design, human resources, operations, and other activities. Rather than the isolated islands of QI, in QIII experiments' results in one area inform decision-making and further experiments in another. As they coordinate multiple experiments across teams and divisions, systematic experimenters embrace both iterative cycles of experimentation and clearly defined protocols of intervention, blending the two logics of experimentation. Yet, as pointed out by Thomke (2020), without the creation of an organizational culture that promotes organization-wide experimentation and tolerates mistakes, reaching such *ideal* might remain a mirage.

[INSERT FIGURE 5 ABOUT HERE]

Interdisciplinary Research

Experimentation in organizations has been studied from various disciplinary lenses including computer science, data science, engineering, entrepreneurship, general management, operations, policy, and economics. What is unique about all these studies is that experimentation provides a range of insights into situations where decision makers rely primarily on data they generate themselves.

Disciplines also differ in their focus. Operations and management studies view the experimentation process as a systematic, iterative capability for navigating uncertainty and learning (e.g., Thomke, 1998). Entrepreneurship and innovation scholars conceptualize experimentation as an engine for opportunity discovery and business model innovation (e.g., Bao et al., 2020). Studies in economics consider experimentation as an investment under uncertainty, treating it as a portfolio of staged bets to discover high-value opportunities in a competitive landscape (e.g., Kerr et al., 2014). Computer science, engineering and operations scholars highlight the importance of causal inference and identification (e.g., Johari et al., 2015). While sharing with other disciplines a dedication to iterative learning under uncertainty, engineering and information technology also places particular emphasis on technical infrastructures and automation as a means to accelerate feedback processes.

Experimentation is recognized in various disciplines as a factor that can influence organizational capability and agility, though its emphasis varies across contexts. Management research mostly looks into this perspective, portraying experimentation as a capability that enables organizations to adapt rapidly and navigate uncertainty (e.g., Bojovic et al., 2018). Entrepreneurship and innovation studies share this adaptive learning focus and particularly emphasize pivoting and flexible processes that help resource-constrained ventures remain agile in uncertain environments (e.g., Khanagha et al., 2014). In contrast, this perspective is modestly addressed in computer science, where experimentation is framed more in terms of short-loop iteration and technical system design. Similarly, economics largely treats experimentation as a mechanism for allocating investment and enabling market-level adjustment.

Digital tools also play different roles across various disciplines. Fields such as computer science, operations, and engineering engage most with this perspective, illustrating how digital infrastructure systems facilitate rapid, scalable, and precise experimentation—including the advancement of "experimentation-as-a-service" (Runge et al., 2020). However, this technological perspective is often decoupled from considerations of organizational routines, politics or strategic alignment. In contrast, management studies recognize the role of digital tools in shortening iteration cycles and enhancing organizational learning but often treat them as one component within broader capability-building frameworks (Waisman et al., 2024). Entrepreneurship and innovation research engages with digital tools and platforms selectively, emphasizing their potential to support business model innovation and ecosystem partnerships rather than their technical properties (Ghezzi, 2019). Economics, for its part, minimally addresses digital tools directly, viewing them mainly as environmental enablers that reduce the cost and risk of experimentation. These distinctions present opportunities to combine the technical depth of engineering with the strategic, organizational, and ecosystem insights found in management and entrepreneurship. Such integration can yield more practical theories regarding how digital tools facilitate rather than drive experimentation within organizations.

Opportunities for Future Research

Our review indicates that each of the two identified logics of experimentation has advanced for the most part in parallel, with limited contributions crossing the two. While the surveyed articles offer important insights on a number of features and characteristics of organizational experiments, opportunities to expand the field abound.

First, because retention failure emerges as a central bottleneck from our analysis, deeper understanding of whether experimentation is followed by *organizational learning* offers an intriguing opportunity for future work. Traditionally, organizational learning has been viewed as deeply tied to experimentation (Huber, 1991; March, 1991). The assumption is that as organizations run experiments, they accumulate direct experience, which translates into learning (Levitt & March, 1988). Yet, our review suggests that organizations sometimes resist learning and at times learn the “wrong thing” from experiments, underscoring that learning is not an automatic outcome. For example, politics and the cognitive distance between those involved in experimentation and those implementing the results can both slow down and prevent effective learning. Besides politics, learning myopia (Levinthal & March, 1993) and organizational forgetting (Rao & Argote, 2006) may also be a factor. An exploration of the organizational factors leading to inert behavior with respect to experimental learning opens exciting avenues for future research, and calls for greater examination of understudied stages (retention) in the VSR framework within each stream.

Our cross-stream analysis also suggests that there is an opportunity to balance the emphasis on speed (C-I stream) with more long-term experimentation (P-S stream). Much research on organizational experiments has highlighted the importance of fast experimentation. Scholars have invoked the “fail fast” concept as a fast track to learning in both nascent and established organizations (Khanna et al., 2016; Koporcic et al., 2024). Practitioners have also advocated for rapid experimentation. Yet, too much emphasis on fast experimentation could lead to poor design choices in how the experiments are structured, thereby incurring the risk of obtaining unsatisfactory or biased results. Speed also often inhibits learning from the results for long-term benefit.

As evidenced in the surveyed literature, not all knowledge generated via experiments is easily codified if not using the appropriate knowledge translation tools (Corbo et al., 2020) that enable the crystallization of knowledge into routines (Nelson & Winter, 1982). In their absence, knowledge may not be able to persist through time in organizations, leading to organizational forgetting. One instance in which knowledge can depreciate over time has been attributed to the departure of individuals from the organization. A high turnover in teams involved in experimentation activities can generate a loss of insights. In other situations, organizational forgetting may be purposeful rather than involuntary. For example, when experiments are run by stable teams over time, individual beliefs may disappear over time and lead to knowledge homogeneity until no further learning occurs (March, 1991). Variation is thus reinserted once team members rotate within or depart from the organization. Therefore, it is relevant for managers of teams in charge of experimentation to understand the timing of intervention to reset the 'learning clock'. Surprisingly, scant research exists linking experimentation and organizational forgetting.

Second, there is an opportunity to expand the *scope of experimentation* within organizations. While many experiments focus on more limited tasks (e.g., pricing models), experimentation on wider scope issues such as strategy or governance (e.g., organizational structure and design choices) is more rare. This is partly because wide-scope experimentation can cause significant changes to the organization, which may not be desirable by its leadership. Experimentation within organizations is influenced by internal politics since for experiments "to travel outside design meetings ... powerful allies in upper management who read them from perspectives of marketing and company profit [are needed]" (Henderson, 1995: 291). This tension opens up intriguing questions for future research. For example, how can organization-wide experiments be used as a source of change and new direction? What experiments can be run on organizational structure? If experiments are influenced by senior decision-makers within an organization (e.g., HiPPO), how can local units leverage personal or informal connections to influence the attention and resources allocated to their experimentation programs? Since business units and local teams are often characterized by self-interest, which actions can senior executives undertake to counterbalance it and promote organization-wide experiments?

Third, there is an interesting opportunity to examine the *tension between science and practice* of experimentation. In organizational experiments, specifying variables so they are well defined and measurable under scientific standards may focus on narrow aspects of the product or phenomenon, and be very different from what the business decision-makers see as complex and dynamic parts of their products and offerings. Similarly, which questions are "worth asking" (from the financial standpoint) plays a huge role in how resources are allocated to experiments. This tension is not new. In the classic work on *"program evaluation"* (Rossi et al., 1999; Campbell, 1969) authors note a tension between (1) "robust" scientific methods of experimentation (to meet research standards) versus (2) the kinds of experiments that organizations actually run (to meet their performance goals). Despite the interest, very few empirical studies have examined this tension in practice. The other point of tension is the time horizon. High-quality research requires time whereas in organizations the time horizon is often managed quarterly. Meeting scientific standards for causality may require such elaborate and esoteric experiments that what is being studied is no longer relevant to the actual business when the results finally arrive. Time horizons become even more fascinating in light of organizations' increasing reliance on AI which contributes to faster experimentation. Overall, ensuring validity for experiments in the scientific sense may contradict ensuring that the findings of the experiment are timely, relevant, and useful for organizations, providing another interesting opportunity for future research.

Conclusion

Organizations have long relied on experiments to guide decision-making but a comprehensive understanding of the literature studying experimentation in organizations remains lacking. We identified two primary streams of research that shape our understanding of experimentation in organizations: (1) problem-solving–based experimentation and (2) causal-inference–driven experimentation. Building on the differences and complementarities between these perspectives, we developed an integrative conceptual framework that highlights key drivers, processes, and outcomes of experimentation in organizational contexts as well as identified the different design choices for each stage of experimentation within the two streams. Our review adds value by providing an integrated synthesis of distinct approaches to experimentation that can contribute to the development of a common “language” to explain experimentation in organizations.

REFERENCES

Alblas, A., & Notten, M. (2021). Speed is significant in short-loop experimental learning: Iterating and debugging in high-tech product innovation. Decision Sciences, 52(6), 1364-1402.

Aldrich, H. (1979). Organizations and environments. Stanford, CA: Stanford University Press.

Aldrich, H. (1999). Organizations evolving. Sage.

Allen, T. J., (1966). Studies of the problem-solving process in engineering design. IEEE Transactions on Engineering Management, (2), 72-83.

Allen, T. J. (1977). Managing the Flow of Technology. MIT Press, Cambridge, MA.

Allen, J. S., Combs, J. G., Carr, J. C., Michaelis, T. L., & Joseph, D. L. (2024). More than one way to pivot: The case for opportunity and survival pivots. Journal of Management, 51(5), 1803-1833.

Allen, R. T., & McDonald, R. M. (2025). Methodological pluralism and innovation in data-driven organizations. Administrative Science Quarterly, 70(2), 403-443.

Andries, P., Debackere, K., & Van Looy, B. (2013). Simultaneous experimentation as a learning strategy: Business model development under uncertainty. Strategic entrepreneurship journal, 7(4), 288-310.

Angus, R. W. (2019). Problemistic search distance and entrepreneurial performance. Strategic Management Journal, 40(12), 2011-2023.

Austin, R. D., Devin, L., & Sullivan, E. E. (2012). Accidental innovation: Supporting valuable unpredictability in the creative process. Organization Science, 23(5), 1505-1522.

Bahçeci, E., Katila, R., & Miikkulainen, R. (2015). Evolving strategies for social innovation games. In Proceedings of the 2015 Annual Conference on Genetic and Evolutionary Computation, 1135-1142.

Bao, Y., Wei, Z., & Di Benedetto, A. (2020). Identifying the tacit entrepreneurial opportunity of latent customer needs in an emerging economy: The effects of experiential market learning versus vicarious market learning. Strategic Entrepreneurship Journal, *14*(3), 444-469.

Becker, S. D., & Endenich, C. (2023). Entrepreneurial Ecosystems as Amplifiers of the Lean Startup Philosophy: Management Control Practices in Earliest-Stage Startups. Contemporary Accounting Research, 40(1), 624-667.

Berends, H., Van Burg, E., & Garud, R. (2021). Pivoting or persevering with venture ideas: Recalibrating temporal commitments. Journal of Business Venturing, 36(4), 106126.

Berg, J. M. (2016). Balancing on the creative highwire: Forecasting the success of novel ideas in organizations. Administrative Science Quarterly, 61(3), 433-468.

Bingham, C. B., & Davis, J. P. (2012). Learning sequences: Their existence, effect, and evolution. Academy of Management Journal, 55(3), 611-641.

Bogers, M., & Horst, W. (2014). Collaborative prototyping: Cross-fertilization of knowledge in prototype-driven problem solving. Journal of Product Innovation Management, 31(4), 744-764.

Bohn, R., & Lapre, M. A. (2011). 11 Accelerated Learning by Experimentation. In Learning curves: Theory, models, and applications (pp. 191-210). CRC Press.

Bojinov I, Saint-Jacques G, Tingley M (2020) Avoid the pitfalls of a/b testing: Make sure your experiments recognize customers' varying needs. Harvard Business Review, 98(2):48–53

Bojinov, I., Simchi-Levi, D., & Zhao, J. (2023). Design and analysis of switchback experiments. Management Science, 69(7), 3759-3777.

Bojovic, N., Genet, C., & Sabatier, V. (2018). Learning, signaling, and convincing: The role of experimentation in the business modeling process. Long Range Planning, 51(1), 141-157.

Bremner, R. P., & Eisenhardt, K. M. (2022). Organizing form, experimentation, and performance: Innovation in the nascent civilian drone industry. Organization Science, 33(4), 1645-1674.
Britan, G. M. (1978). Experimental and contextual models of program evaluation. Evaluation and Program Planning, 1(3), 229-234.
Browder, R. E., Koch, H., Long, A., & Hernandez, J. M. (2022). Learning to innovate with big data analytics in interorganizational relationships. Academy of Management Discoveries, 8(1), 139-166.
Brown, S. L., & Eisenhardt, K. M. (1997). The art of continuous change. Administrative Science Quarterly, 42(1), 1-34.
Bruneel, J., Yli-Renko, H., & Clarysse, B. (2010). Learning from experience and learning from others: How congenital and interorganizational learning substitute for experiential learning in young firm internationalization. Strategic entrepreneurship journal, 4(2), 164-182.
Burgelman, R. A. (1983). A process model of internal corporate venturing in the diversified major firm. Administrative Science Quarterly, 28(2), 223–244.
Burnell, D., Stevenson, R., & Fisher, G. (2023). Early-stage business model experimentation and pivoting. Journal of Business Venturing, 38(4), 106314.
Burnell, D., Fisher, G., Stevenson, R., & Kuratko, D. F. (2025). Entrepreneurial Experimentation: Conceptual Foundations, Integrative Theoretical Framework, and Research Agenda. Entrepreneurship Theory and Practice, 10422587251347046.
Campbell, D. T. (1969). Reforms as experiments. American Psychologist, 24(4), 409.
Camuffo, A., Cordova, A., Gambardella, A., & Spina, C. (2020). A scientific approach to entrepreneurial decision making: Evidence from a randomized control trial. Management Science, 66(2), 564–586.
Camuffo, A., Gambardella, A., Messinese, D., Novelli, E., Paolucci, E., & Spina, C. (2024). A scientific approach to entrepreneurial decision-making: Large-scale replication and extension. Strategic Management Journal, 45(6), 1209-1237.
Chamandy, N. (2016) Experimentation in a ridesharing marketplace—Lyft engineering, available at https://eng.lyft.com/experimentation-in-a-ridesharing-marketplace-b39db027a66e.
Chen, E. L., & Katila, R. (2008). Rival interpretations of balancing exploration and exploitation: Simultaneous or sequential. In S. Shane (Ed.), Handbook of technology and innovation management (pp. 197–214). Wiley.
Chen, X., & Zhang, S. X. (2024). Too much of two good things: The curvilinear effects of self-efficacy and market validation in new ventures. Journal of Business Research, 183, 114845.
Contigiani, A., & Young-Hyman, T. (2022). Experimentation, planning, and structure in early-stage ventures: Evidence from pitch decks. Strategic Entrepreneurship Journal, 16(3), 425-459.
Cook, T. D., Campbell, D. T. (1979). Quasi-experimentation: Design & analysis issues for field settings. Boston: Houghton Mifflin.
Cooprider, J., & Nassiri, S. (2023). Science of price experimentation at Amazon. *Business Economics*, *58*(1), 34-41.
Corbo, L., Mahassel, S., & Ferraris, A. (2020). Translational mechanisms in business model design: introducing the continuous validation framework. Management Decision, 58(9), 2011-2026.
Cui, A. S., & Wu, F. (2017). The impact of customer involvement on new product development: Contingent and substitutive effects. Journal of Product Innovation management, 34(1), 60-80.
Cyert, R. M., & March, J. G. (1963). A behavioral theory of the firm. Prentice Hall/Pearson Education.

Dalpiaz, E., Rindova, V., & Ravasi, D. (2016). Combining logics to transform organizational agency: Blending industry and art at Alessi. Administrative science quarterly, 61(3), 347-392.
Eisenhardt, K. M. (1989). Making fast strategic decisions in high-velocity environments. Academy of Management Journal, 32(3), 543-576.
Ejdemyr, S., Tingley, M., Shang, Y., & Brooks, T. (2024). Estimating the returns from an experimentation program. In ACIC Conference.
Feng, E. Zhao, Z. (2017) Building an Intelligent Experimentation Platform with Uber Engineering. Available at https://eng.uber.com/experimentationplatform/.
Fisher, R. (1935/1949). The design of experiments. Oliver and Boyd, London
Fligstein, N., & McAdam, D. (2011). Toward a general theory of strategic action fields. Sociological Theory, 29(1), 1-26.
Garud R., Karunakaran A. (2018). Enabling identity-challenging innovations through participative experimentation: The case of Gmail and Adsense at Google. *Strategic Organization*, 16, 273–303.
Ghezzi, A. (2019). Digital startups and the adoption and implementation of Lean Startup Approaches: Effectuation, Bricolage and Opportunity Creation in practice. Technological Forecasting and Social Change, 146, 945-960.
Gillier, T., & Lenfle, S. (2019). Experimenting in the unknown: lessons from the Manhattan project. European Management Review, 16(2), 449-469.
Glynn, M. A., & D'Aunno, T. (2023). An intellectual history of institutional theory: Looking back to move forward. Academy of Management Annals, 17(1), 301-330.
Grimes, M. G. (2018). The pivot: How founders respond to feedback through idea and identity work. Academy of Management Journal, 61(5), 1692-1717.
Hacking, I. (1983). Representing and intervening: Introductory topics in the philosophy of natural science. Cambridge University Press.
Harms, R., & Schwery, M. (2020). Lean startup: operationalizing lean startup capability and testing its performance implications. Journal of small business management, 58(1), 200-223.
Harvey, J. F., Cromwell, J. R., Johnson, K. J., & Edmondson, A. C. (2023). The dynamics of team learning: Harmony and rhythm in teamwork arrangements for innovation. Administrative Science Quarterly, 68(3), 601-647.
Hauptman, O., & Iwaki, G. (1991). The final voyage of the Challenger. Harvard Business School case.
Hauser, O., Linos, E. & Rogers, T. (2017). Innovation with field experiments: Studying organizational behaviors in actual organizations, Research in Organizational Behavior, 185-198.
Helfat, C. E., & Peteraf, M. A. (2015). Managerial cognitive capabilities and the microfoundations of dynamic capabilities. Strategic Management Journal, 36(6), 831-850.
Helfat, C. E., Finkelstein, S., Mitchell, W., Peteraf, M., Singh, H., Teece, D., & Winter, S. G. (2009). Dynamic capabilities: Understanding strategic change in organizations. Wiley-Blackwell.
Henderson, K. (1995). The political career of a prototype: Visual representation in design engineering. Social Problems, 42(2), 274-299.
Holtz, D., Lobel, R., Liskovich, I., & Aral, S. (2020). Reducing interference bias in online marketplace pricing experiments. Management Science, 66(9), 4030–4048.
Huber, G. P. (1991). Organizational learning: The contributing processes and the literatures. Organization Science, 2(1), 88-115.
Johari, R., Pekelis, L., & Walsh, D. J. (2015). Always valid inference: Bringing sequential analysis to A/B testing. arXiv preprint arXiv:1512.04922.

Kalvapalle, S. G., Phillips, N., & Cornelissen, J. (2024). Entrepreneurial pitching: A critical review and integrative framework. Academy of Management Annals, 18(2), 550-599.
Katila, R., & Ahuja, G. (2002). Something old, something new: A longitudinal study of search behavior and new product introduction. Academy of Management Journal, 45(6), 1183-1194.
Katila, R., Thatchenkery, S., Christensen, M. Q., & Zenios, S. (2017). Is there a doctor in the house? Expert product users, organizational roles, and innovation. Academy of Management Journal, 60(6), 2415-2437.
Kerr, W. R., Nanda, R., & Rhodes-Kropf, M. (2014). Entrepreneurship as experimentation. Journal of Economic Perspectives, 28(3), 25-48.
Khanagha, S., Volberda, H., & Oshri, I. (2014). Business model renewal and ambidexterity: structural alteration and strategy formation process during transition to a C loud business model. R&D Management, 44(3), 322-340.
Khanna, R., Guler, I., & Nerkar, A. (2016). Fail often, fail big, and fail fast? Learning from small failures and R&D performance in the pharmaceutical industry. Academy of Management Journal, 59(2), 436-459.
Kim, H., & Luca, M. (2019). Product quality and entering through tying: Experimental evidence. Management Science, 65(2), 596-603.
Kirtley, J., & O'Mahony, S. (2023). What is a pivot? Explaining when and how entrepreneurial firms decide to make strategic change and pivot. Strategic Management Journal, 44(1), 197-230.
Kohavi, R., Crook, T., Longbotham, R., Frasca, B., Henne, R., Ferres, JL, & Melamed, T. (2009) Online experimentation at Microsoft. 3rd Data Mining Case Studies workshop 11:39.
Kohavi, R. (2010). Online controlled experiments: Listening to the customers, not to the HiPPO. Keynote at EC10 AC Conference on Electronic Commerce.
Kohavi, R., Deng, A., Frasca, B., Walker, T., Xu, Y., & Pohlmann, N. (2013). Online controlled experiments at large scale. In Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 1168-1176).
Kohavi, R., & Thomke, S. (2017). The surprising power of online experiments. *Harvard Business Review*, *95*(5), 74-82.
Kohavi, R., Tang, D., & Xu, Y. (2020). Trustworthy online controlled experiments: A practical guide to A/B testing. Cambridge University Press.
Konietzko, J., Baldassarre, B., Brown, P., Bocken, N., & Hultink, E. J. (2020). Circular business model experimentation: Demystifying assumptions. Journal of Cleaner Production, 277, 122596.
Koning, R., Hasan, S., & Chatterji, A. (2022). Experimentation and start-up performance: Evidence from A/B testing. Management Science, 68(9), 6434-6453.
Koporcic, N., Sjödin, D., Kohtamäki, M., & Parida, V. (2024). Embracing the "fail fast and learn fast" mindset: conceptualizing learning from failure in knowledge-intensive SMEs. Small Business Economics, 64(1), 181-202.
Kossek, E. E., & Perrigino, M. B. (2016). Resilience: A review using a grounded integrated occupational approach. Academy of Management Annals, 10(1), 729-797.
Larsen, B. (2002). Exploiting citation overlaps for information retrieval: generating a boomerang effect from the network of scientific papers. Scientometrics, 54, 155-178.
Larsen, N. Stallrich, J. Sengupta, A. Deng, A., Kohavi R. & Stevens, N. (2024) Statistical Challenges in Online Controlled Experiments: A Review of A/B Testing Methodology, The American Statistician, 78:2, 135-149.

Leatherbee, M., & Katila, R. (2020). The lean startup method: Early-stage teams and hypothesis-based probing of business ideas. Strategic Entrepreneurship Journal, 14(4), 570-593.

Lee, F., Edmondson, A. C., Thomke, S., & Worline, M. (2004). The mixed effects of inconsistency on experimentation in organizations. Organization Science, 15(3), 310-326.

Levinthal, D. A., & March, J. G. (1993). The myopia of learning. Strategic management journal, 14(S2), 95-112.

Levinthal, D. A. (2017). Mendel in the C-suite: Design and the evolution of strategies. Strategy Science, 2(4), 282-287.

Levitt, B., & March, J. G. (1988). Organizational learning. Annual review of sociology, 14(1), 319-338.

Levitt, S., & List, J. (2009). Was there Really a Hawthorne Effect at the Hawthorne Plant? An Analysis of the Original Illumination Experiments. NBER Working Paper No. 15016.

Loch, C. H., Terwiesch, C., & Thomke, S. (2001). Parallel and sequential testing of design alternatives. Management Science, 47(5), 663-678.

Lukkarinen, A. & Katila, R. 2026. When Sustainability Labels Backfire: A Randomized Field Experiment with Equity Investors. Stanford Technology Ventures Program working paper.

Lynn, G. S., Morone, J. G., & Paulson, A. S. (1996). Marketing and discontinuous innovation: the probe and learn process. California Management Review, 38(3), 8-37.

Mansoori, Y., Karlsson, T., & Lundqvist, M. (2019). The influence of the lean startup methodology on entrepreneur-coach relationships in the context of a startup accelerator. Technovation, 84, 37-47.

March, J. G. (1981). Footnotes to organizational change. Administrative Science Quarterly, 563-577.

March, J. G. (1991). Exploration and exploitation in organizational learning. Organization Science, 2(1), 71-87.

Marres, N., & Stark, D. (2020). Put to the test: For a new sociology of testing. The British Journal of Sociology, 71(3), 423-443.

McDonald, R., & Gao, C. (2019). Pivoting isn't enough? Managing strategic reorientation in new ventures. Organization Science, 30(6), 1289-1318.

McGrath, R. G. (2024). Who learns fastest, wins: lean startup and discovery driven growth. Journal of Management, 50(8), 3162-3182.

Millet, F. C., Lewis, S., & Stoddart, P. (2022). Local Search Quality: Comment on "Product Quality and Entering Through Tying: Experimental Evidence". Management Science, 68(4), 3169-3174.

Miner, A. S., Haunschild, P. R., & Schwab, A. (2003). Experience and convergence: Curiosities and speculation. Industrial and Corporate Change, 12(4), 789-813.

Motley, C., Leatherbee, M., & Katila, R. (2025). From critique to catalyst: How academic entrepreneurs transform negative feedback into pivots and performance. Strategic Entrepreneurship Journal.

Murray, F., & Tripsas, M. (2004). The exploratory processes of entrepreneurial firms: The role of purposeful experimentation. In Business strategy over the industry lifecycle (pp. 45-75). Emerald Group Publishing Limited.

Na, C., Khalid, R., Raza, M., Ramirez-Asis, E., Huerta-Soto, R., & Jahanger, A. (2022). The role of age and business coaching in the relationship of lean startup approach and innovative work behavior of women entrepreneurs during COVID-19. Frontiers in Psychology, 13, 946918.

Nelson, R. R., & Winter, S. G. (1982). An evolutionary theory of economic change. Cambridge, MA: Harvard University Press.

Nicholas, J., Ledwith, A., & Bessant, J. (2015). Selecting early-stage ideas for radical innovation: Tools and structures. Research-Technology Management, 58(4), 36-44.

Nicholls-Nixon, C. L., Cooper, A. C., & Woo, C. Y. (2000). Strategic experimentation: Understanding change and performance in new ventures. Journal of Business Venturing, 15(5-6), 493-521.

Orlikowski, W. J. (1996). Improvising organizational transformation over time: A situated change perspective. Information systems research, 7(1), 63-92.

Ott, T. E., Eisenhardt, K. M., & Bingham, C. B. (2017). Strategy formation in entrepreneurial settings: Past insights and future directions. Strategic Entrepreneurship Journal, 11(3), 306-325.

Ott, T. E., Eisenhardt, K. M., & Bingham, C. B. (2017). Strategy formation in entrepreneurial settings: Past insights and future directions. Strategic Entrepreneurship Journal, 11(3), 306-325.

Paiola, M., Agostini, L., Grandinetti, R., & Nosella, A. (2022). The process of business model innovation driven by IoT: Exploring the case of incumbent SMEs. Industrial Marketing Management, 103, 30-46.

Paust, S., Korsgaard, S., & Thrane, C. (2024). Entrepreneurial prototyping: the role of purpose, prototype recycling, and skills bricolage. Small Business Economics, 1-25.

Peprah, A. A., Giachetti, C., Larsen, M. M., & Rajwani, T. S. (2022). How business models evolve in weak institutional environments: The case of Jumia, the Amazon. com of Africa. Organization Science, 33(1), 431-463.

Phillips, R. (2025). Data science and experiments [Book manuscript in progress].

Pillai, S. D., Goldfarb, B., & Kirsch, D. A. (2020). The origins of firm strategy: Learning by economic experimentation and strategic pivots in the early automobile industry. Strategic Management Journal, 41(3), 369-399.

Posen, H. E., & Chen, J. S. (2013). An advantage of newness: Vicarious learning despite limited absorptive capacity. Organization Science, 24(6), 1701-1716.

Posen, H. E., & Levinthal, D. A. (2012). Chasing a moving target: Exploitation and exploration in dynamic environments. Management science, 58(3), 587-601.

Rahman, H. (2024). Inside the invisible cage: How algorithms control workers. University of California Press.

Raneri, S., Lecron, F., Hermans, J., & Fouss, F. (2023). Predictions through Lean startup? Harnessing AI-based predictions under uncertainty. International Journal of Entrepreneurial Behavior & Research, 29(4), 886-912.

Rao, R. D., & Argote, L. (2006). Organizational learning and forgetting: The effects of turnover and structure. European Management Review, *3*(2), 77-85. “

Rathje, J., Katila, R., & Reineke, P. (2024). Making the most of AI and machine learning in organizations and strategy research: Supervised machine learning, causal inference, and matching models. Strategic Management Journal, 45(10), 1926-1953.

Rathje, J., Katila, R., & Reineke, P. (2024). Making the most of AI and machine learning in organizations and strategy research: Supervised machine learning, causal inference, and matching models. Strategic Management Journal, 45(10), 1926-1953.

Reineke, P., & Katila, R. 2025. Shadow decentralization in community organizations. *Working paper.*

Reineke, P., Katila, R., & Eisenhardt, K. M. (2025). Decentralization in Organizations: A Revolution or a Mirage?. Academy of Management Annals, 19(1), 298-342.

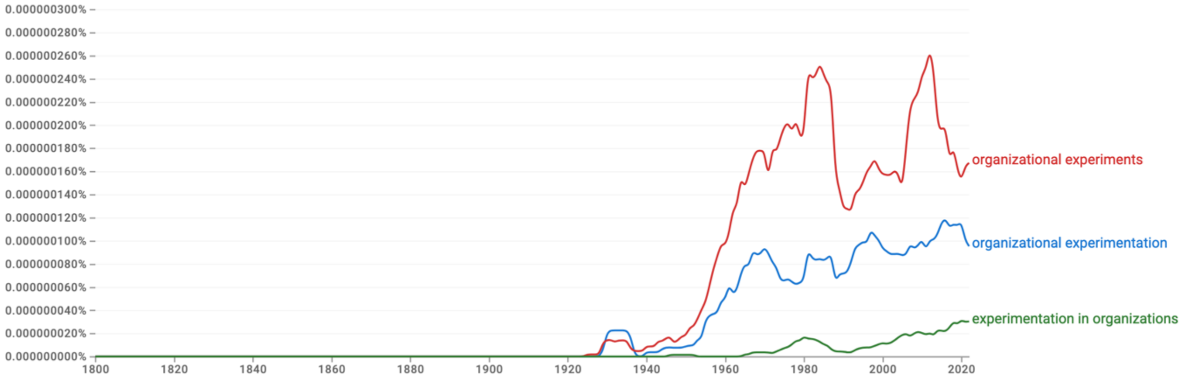
Rosenberg, N. (1982). Inside the Black Box. Cambridge University Press.

Rossi, Peter H., Freeman, Howard E., and Lipsey, Mark W. (1999). Evaluation. A Systematic Approach (6th ed.). Thousand Oaks, CA: Sage

Rousseau, D. M. (2024). Reviews as research: Steps in developing trustworthy synthesis. Academy of Management Annals, 18(2), 395-402.
Runge, J., Geinitz, S., & Ejdemyr, S. (2020). Experimentation and performance in advertising: An observational survey of firm practices on Facebook. Expert Systems with Applications, 158, 113554.
Sanasi, S., Manotti, J., & Ghezzi, A. (2022). Achieving agility in high-reputation firms: Agile experimentation revisited. IEEE Transactions on Engineering Management, 69(6), 3529-3545.
Sarta, A., Durand, R., & Vergne, J. P. (2021). Organizational adaptation. Journal of Management, 47(1), 43-75.
Schmickl, C., & Kieser, A. (2008). How much do specialists have to learn from each other when they jointly develop radical product innovations?. Research Policy, 37(3), 473-491.
Shepherd, D. A., & Gruber, M. (2021). The lean startup framework: Closing the academic–practitioner divide. Entrepreneurship theory and practice, 45(5), 967-998.
Sitkin, S. B. (1992). Learning through failure: The strategy of small losses. Research in Organizational Behavior, 14, 231-266.
Snihur, Y., & Clarysse, B. (2022). Sowing the seeds of failure: Organizational identity dynamics in new venture pivoting. Journal of Business Venturing, 37(1), 106164.
Solaimani, S., van Eck, T., Kievit, H., & Koelemeijer, K. (2022). An exploration of the applicability of Lean Startup in small non-digital firms: an effectuation perspective. International Journal of Entrepreneurial Behavior & Research, 28(9), 198-218.
Sorenson, O. (2003). Strategy as quasi-experimentation. Strategic Organization, 1(3), 337-343.
Staw, B. M. (1977). The experimenting organization. Organizational dynamics, 6(1), 3-18.
Stevenson, R., Josefy, M., McMullen, J. S., & Shepherd, D. (2020). Organizational and management theorizing using experiment-based entrepreneurship research: Covered terrain and new frontiers. Academy of Management Annals, 14(2), 759-796.
Stevenson, R., Burnell, D., & Fisher, G. (2024). The Minimum Viable Product (MVP): Theory and Practice. Journal of Management, *50*(8), 3202-3231.
Suchman, E. (1968). Evaluative research: Principles and practice in public service and social action programs. Russell Sage Foundation, New York.
Terwiesch, C., & Loch, C. H. (2004). Collaborative prototyping and the pricing of custom-designed products. Management Science, 50(2), 145-158.
Terwiesch, C., & Ulrich, K. (2009). Innovation tournaments: Creating and selecting exceptional opportunities. Harvard Business Press.
Thatchenkery, S., & Katila, R. (2021). Seeing what others miss: A competition network lens on product innovation. Organization Science, 32(5), 1346–1370.
Thatchenkery, S., & Katila, R. (2023). Innovation and profitability following antitrust intervention against a dominant platform: The wild, wild west? Strategic Management Journal, 44(4), 943–976.
Thau, S., Pitesa, M., & Pillutla, M. (2014). Experiments in organizational behavior. In M. Webster & J. Sell (Eds.), Laboratory experiments in the social sciences (2nd ed., pp. 285–310). Elsevier.
Thomke, S. H. (1998). Managing experimentation in the design of new products. Management Science, 44(6), 743-762.
Thomke, S., & Fujimoto, T. (2000). The effect of "front-loading" problem-solving on product development performance. Journal of Product Innovation Management, 17(2), 128-142.
Thomke, S., & Bell, D. E. (2001). Sequential testing in product development. Management Science, 47(2), 308-323.
Thomke, S. H. (2020). Building a culture of experimentation. Harvard Business Review, 98(2), 40-47.

Thomke, S. H., & Loveman, G. (2022). Act like a scientist. Harvard Business Review, 100(2), 48–57.
Thompson, P. (2010). Learning by doing. In Handbook of the Economics of Innovation, 1, 429-476.
Valentine, J., Novelli, E., & Agarwal, R. (2024). The theory-based view and strategic pivots: The effects of theorization and experimentation on the type and nature of pivots. Strategy Science, 9(4), 433-460.
Waisman, C., Sahni, N. S., Nair, H. S., & Lin, X. (2025). Parallel experimentation and competitive interference on online advertising platforms. Marketing Science, *44*(2), 437-456.
Weissbrod, I., & Bocken, N. M. (2017). Developing sustainable business experimentation capability–A case study. Journal of Cleaner Production, 142, 2663-2676.
West, J., & Iansiti, M. (2003). Experience, experimentation, and the accumulation of knowledge: the evolution of R&D in the semiconductor industry. Research Policy, 32(5), 809-825.
Woodward, J. (2005). Making things happen: A theory of causal explanation. Oxford University Press.
Wu, L., Lou, B., & Hitt, L. (2019). Data analytics supports decentralized innovation. Management Science, 65(10), 4863-4877.
Yaqub, O. (2018). Variation in the dynamics and performance of industrial innovation: What can we learn from vaccines and HIV vaccines?. Industrial and Corporate Change, 27(1), 173-187.
Zhao, Y., Li, Y., Lee, S. H., & Chen, L. B. (2011). Entrepreneurial orientation, organizational learning, and performance: Evidence from China. Entrepreneurship Theory and Practice, 35(2), 293–317.
Zhuge, K., He, H., Yuan, Y., & Sun, P. (2023). Can adopting lean startup strategy promote the sustainable development of new ventures? The mediating role of organizational iterative learning. Plos one, 18(8), e0290849.
Zou, T., Xiong, F., Li, S., & Zhang, W. (2025). Understanding the Determinants of Firms' Usage of A/B Testing: A Technology-Organization-Environment Framework. IEEE Transactions on Engineering Management, 72, 378-400.
Zuzul, T., & Tripsas, M. (2020). Start-up inertia versus flexibility: The role of founder identity in a nascent industry. Administrative Science Quarterly, 65(2), 395-433.

FIGURE 1. Frequency of terms related to organizational experimentation



Source: Google Ngram Viewer.

FIGURE 2. Sampling Protocol

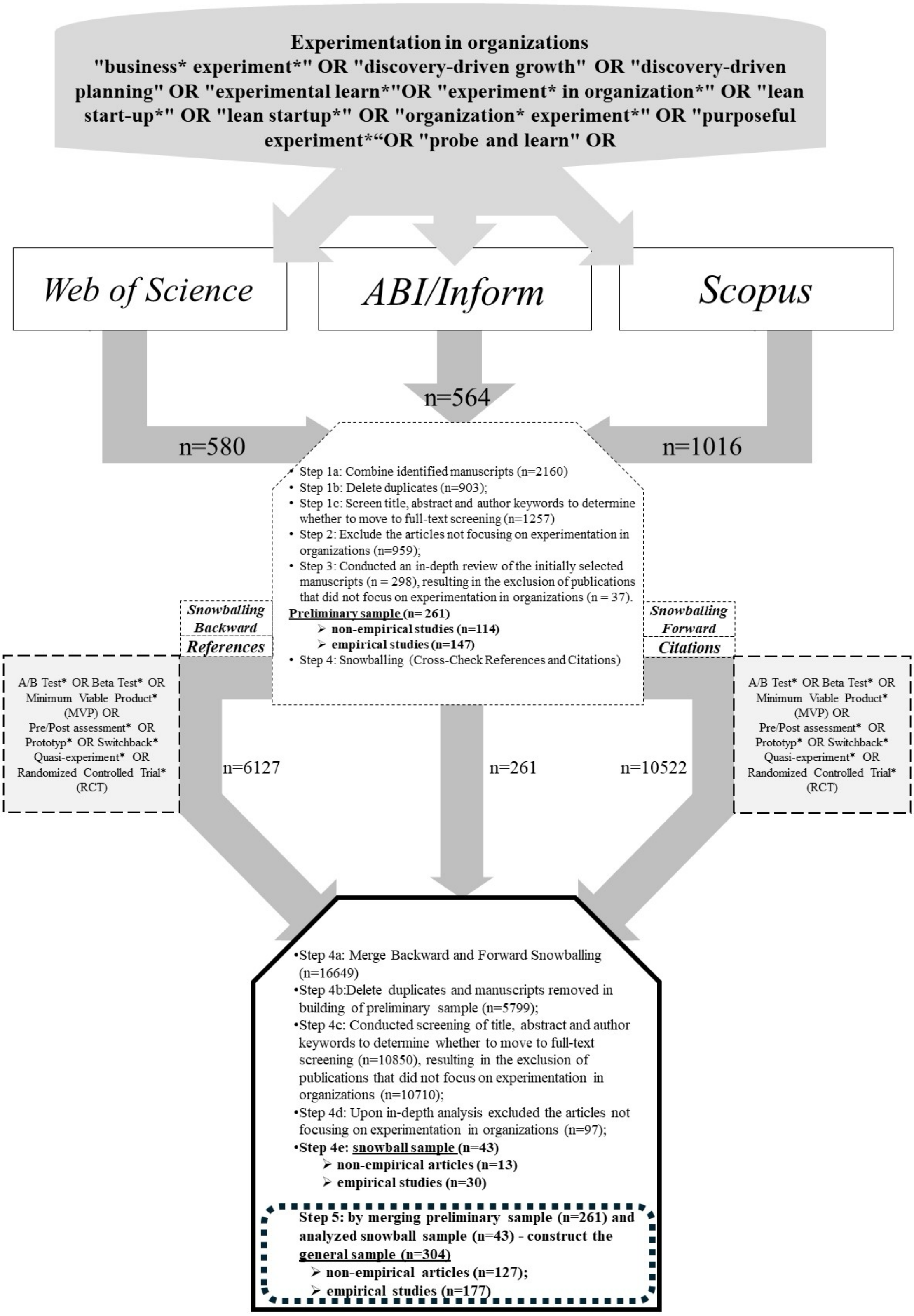


TABLE 1. Distribution of Publications Across Fields

Management	Entrepreneurship and Innovation	Computer Science, Operations & Engineering	Other
41%	**33%**	**18%**	**8%**
Journals	**Journals**	**Journals**	**Journals**
Strategic Management Journal	International Journal of Entrepreneurial Behaviour and Research	IEEE Transactions on Engineering Management	Journal of Cleaner Production
Organization Science	Strategic Entrepreneurship Journal	Journal of Systems and Software	Frontiers in Psychology
Administrative Science Quarterly	Journal of Business Venturing	Information and Software Technology	Business Economics
Management Science	Research Policy	Decision Science	Economics and Policy of Energy and the Environment
Academy of Management Journal	Journal of Product Innovation Management	Information Systems Research	Journal of Workplace Learning

Note: Analysis conducted on the empirical papers (N=177). Full list of empirical manuscripts available in Appendix A. For a similar categorization of the literature across disciplines, see Kalvapalle et al. (2024).

TABLE 2. Exemplary Design Choices in Experimentation

Design Choice	P-S examples	C-I examples
A/B Testing		• Mud Jeans tested two propositions for leasing jeans in an A/B split test on Facebook...Two advertisements for leasing jeans were placed on Facebook with the same overall target group. The image was kept the same − only the text was changed to test which type of focus would appeal more to the target group. The experiment showed that customers preferred another direction than was initially expected (quoted in Bocken et al., 2018).
Beta Testing	• We did a lot of things wrong during the 2.5 years of pre-launch Gmail development, but one thing we did very right was to always have live code … Literally from day one, we had users internally. One nice thing about Google is that we can just release things internally and have this great population of testers, essentially (quoted in Garud & Karunakaran, 2018).	
Minimum Viable Product (MVP)	• Houston developed an explainer-video MVP to test market interest in a potential product that could allow for seamless file syncing across devices. Rather than first build the full product, which would require multifaceted integration across operating systems, management of large files over slow internet connections, and handling of file conflicts, Houston simply created a video that demonstrated the value of the potential software to gauge interest. The MVP, although nonfunctional, helped Houston ensure that people were interested in the product and resulted in a waiting list of 75,000 people for the beta product within days (quoted in Stevenson, Burnell & Fisher, 2024).	
Prototyping	• In order to reduce risk, the bank's product development team constructed a full-scale foam prototype of a proposed new branch configuration for internal testing. The bank then implemented the new configuration in part of a single branch to test it with customers before finally implementing it across its network (quoted in Nicholas, Ledwith, & Bessant, 2015).	

Design Choice	P-S examples	C-I examples
	• In 2012, Ohm added a new product to their strategy after identifying an opportunity trigger: an internal development project revealed the potential for a new product not previously considered. […] Ohm's engineers previously did not think this was possible as microchips are typically orders of magnitude smaller and faster than discretes. CEO Cam Fahey described how the new prototype built from discretes triggered the team to question whether to change their strategy and develop a second commercial product based on discretes rather than a microchip design: Al said, “I think we might be able to take it to production in discretes.” The only reason he was able to be pretty confident about that is because we had built something in discretes in the form factor. We are trying to squeeze everything into the little, tiny space. Prior to having this [working prototype], there were people on the team who thought it was impossible to make it out of discretes in the form factor (quoted in Kirtley & O´Mahony, 2023).	
Switchbacks		• In order to improve prices, Amazon created Pricing Labs, a price experimentation platform. […] In essence, Amazon repeatedly cycles a product between treatment and a control condition over time. For the time-bound cases, products would switch between the new machine learning (ML) algorithm price and the old ML algorithm price multiple times. […] In this setting, since the treatments start on different days for different products, it allows separating demand shifts across Amazon or among groups of products from the treatment more effectively (quoted in Cooprider & Nassiri, 2023).

TABLE 3. Dimensions of variation-selection-retention across the P-S and C-I logics of experimentation

Dimension	Problem-Solving	Causal Inference
1. Purpose	Novel solutions to ill-structured problems	Tightly-scoped causal effects
2. Variation	Emergent, experience-driven, sometimes ad-hoc; bundled	Intentionally designed, pre-specified, structured and systematic; isolated
3. Selection	Somewhat subjective, interpretive, often identity-filtered	Statistically significant lift or effect sizes
4. Retention	Routines, norms, or identity coherence; in practice, often partial or local	Code, dashboards, platform algorithms; in practice, often incomplete 'tribal memory'
5. Role of Failure	Learning triggers that spark new problem-solving cycles	Undesired and minimized through design
6. Role of Theory	Often under-theorized; grounded in local context and experience; abductive	Explicitly hypothesis-driven, but not theorized; deductive
7. Temporal Horizon	Medium and sometimes long-term	Short-term
8. Design Choices	System-level: MVPs, prototypes, simulations	Algorithm-level: A/B tests, RCTs, switchbacks
9. Level of Formalization	Often informal and context-specific; little infrastructure for codifying experimentation	Often highly formalized; relies on statistica validity, control groups, and experiment infrastructure
10. Validity Trade-offs	Relative emphasis on external and conclusion validity	Relative emphasis on internal and statistical conclusion validity

FIGURE 3. Design choices across experimentation stages

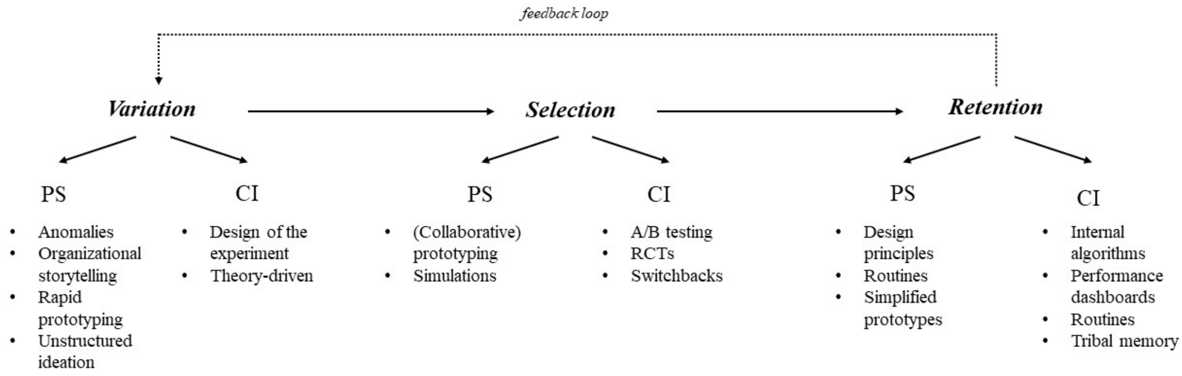


FIGURE 4. Integrative Conceptual Framework

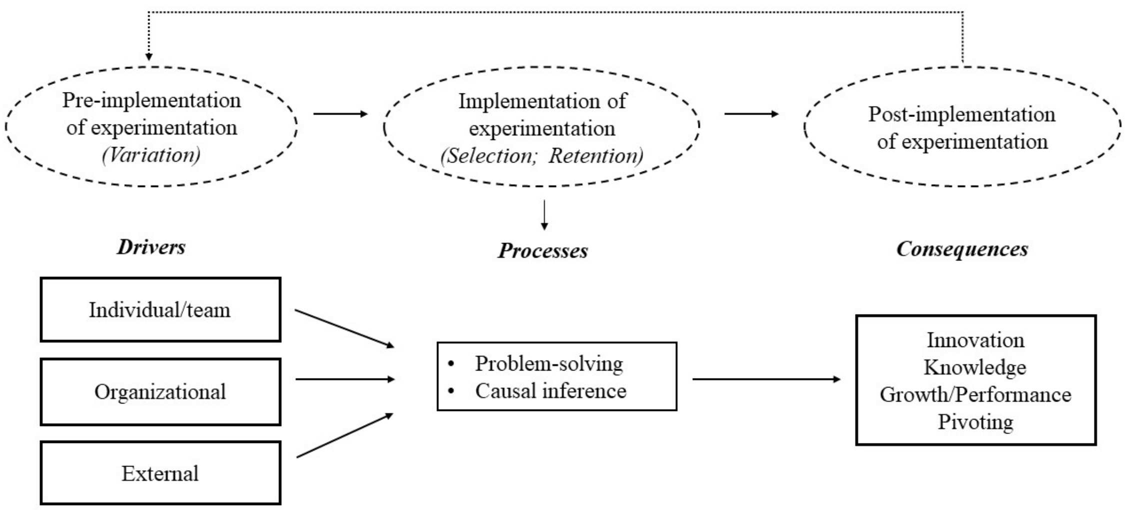


FIGURE 5. Typologies of Experimenting Organizations