Cognitive Maps and Exploratory Search:

When Deviating from Plans Generates Novelty

Stijn Kelchtermans[1,2]

Cindy Lopes-Bento[1,4]

Andrés Madariaga Espinoza[1]

Arvids A. Ziedonis[1,3]

March 24, 2026

ABSTRACT

Organizations often initiate innovation through plans that articulate initial problem formulations, yet novel outcomes frequently emerge when projects deviate from these plans. We examine when such deviation functions as productive exploratory search. Conceptualizing initial problem formulation as a cognitive map, we argue that the value of deviation depends on the coherence of the map with prior knowledge and on the scope of knowledge recombination introduced through team changes. Using a longitudinal dataset that links project formulations to subsequent innovative outputs, we demonstrate that deviation is positively associated with novelty once endogeneity is controlled for. This relationship is significantly stronger when the initial formulations is coherent, or anchored in prior knowledge and expertise. Moreover, within coherent projects, deviation generates greater novelty when it is supported by distant rather than local knowledge recombination through new collaborators. These findings advance strategy and innovation research by clarifying how planning structures exploratory search and by identifying when deviation from initial plans enhances novel outcomes.

Keywords: *Problem formulation; Exploratory search; Knowledge recombination; Novelty*

Lopes-Bento acknowledges funding from the Luxembourg National Research Fund (FNR)

1 K.U. Leuven

2 Hasselt University

4 FNR (Luxembourg National Research Fund)

3 Boston University

INTRODUCTION

Innovative activity often includes an emphasis on planning, ranging from the development of project proposals and roadmaps to investment agendas. Organizational plans can define problems, allocate resources, and guide action. Despite the central role of planning in innovation, however, prior research has long emphasized that novelty rarely emerges from the execution of plans as envisioned at the outset. Novel outcomes instead often emerge when actors adapt, deviate, and explore beyond initial formulations as new information becomes available (March, 1991; Katila & Ahuja, 2002; Thomke, 1997; Lenfle & Loch, 2010). This tension raises a fundamental question for strategy and innovation: when does deviation from initial plans facilitate exploratory search that leads to novel outcomes?

The existing literature provides partial answers. Research on search and problem formulation highlights that structured representations of problems enable actors to navigate complex spaces by constraining attention and guiding experimentation (Gavetti & Levinthal, 2000; Nickerson & Zenger, 2004). At the same time, work on creativity and knowledge recombination emphasizes the value of departing from established paths and integrating diverse knowledge elements to generate novelty (Fleming, 2001; Perry-Smith & Mannucci, 2017). However, these literatures have rarely been integrated. As a result, we lack a clear understanding of when deviation from initial plans improves outcomes from exploration. We address this gap by conceptualizing initial plans as cognitive maps that structure the search space and by theorizing how deviation operates within that space. We argue that the payoff to deviation depends on two interrelated factors. First, the *coherence* of the initial problem formulation, or its alignment with the problem solver's prior expertise, shapes whether exploratory outcomes can be integrated with prior knowledge. Second, the *scope of knowledge recombination* introduced through the introduction of new team members determines the degree to which deviation expands the search space in fruitful ways. Together, these factors imply that deviation is most likely to generate novelty when it occurs within a coherent search space and when it draws on distant knowledge that can be meaningfully integrated. Absent such coherence, distant inputs are harder to absorb and deviation is less likely to successfully recombine knowledge into novel outcomes.

We test these arguments using a longitudinal dataset that links initial project formulations to changes in team composition and subsequent outputs. Leveraging natural language processing techniques, we measure *deviation* as the semantic distance between plans and outcomes, *coherence* as the anchoring of plans in prior knowledge, and *recombination scope* as the distance between new collaborators' prior work and the original problem space. Our findings show that deviation from initial plans is not uniformly beneficial. Instead, deviation from initial formulation is more likely to be successful when coherent and when exploratory search incorporates distant knowledge through new collaborators.

Our study contributes to strategy and innovation research in three ways. First, it reframes the role of planning in innovation by demonstrating how it structures exploratory search. Second, it clarifies when deviation from initial plans enhances novelty by identifying coherence or alignment with prior knowledge and recombination scope as critical boundary conditions. Third, it bridges research on problem formulation, search, and team-based knowledge recombination, offering a more integrated account of how novelty emerges in complex innovative efforts. Together, these insights advance our understanding of how organizations balance structure and flexibility in the pursuit of innovation.

THEORY AND HYPOTHESES

Initial problem formulations as cognitive maps for search

Innovative efforts typically begin with an initial articulation of the problem to be addressed, often formalized in proposals, roadmaps, or project plans. Beyond their administrative role, such formulations shape how actors understand the problem space, guide subsequent search (Gavetti & Levinthal, 2000; Nickerson & Zenger, 2004), and allocate resources. We conceptualize these formulations as cognitive maps or structured representations that locate a plan within the idea generator's existing knowledge-base, define salient dimensions of the problem, and delimit plausible directions for exploration.

Well-articulated problem formulations help actors improve the navigation of complex search spaces by providing heuristics for evaluating new information and integrating emerging insights (Kaplan & Vakili,

2015). At the same time, innovation rarely unfolds exactly as planned. As innovative efforts progress, actors encounter unexpected results, technical obstacles, and new opportunities that prompt adjustments to the initial formulation. Understanding how such adjustments affect outcomes requires a theory of deviation within search, rather than viewing deviation simply as a departure from initial plans.

Deviation as exploratory search

Deviation from an initial problem formulation reflects the extent to which subsequent effort departs from the original cognitive map. From a search perspective, deviation captures exploratory moves that reorient attention, introduce new solution paths, or reinterpret the original problem (March, 1991). Prior work on creativity and innovation suggests that such exploration is a key source of novelty, as it enables the recombination of knowledge elements in new ways (Fleming, 2001; Perry-Smith & Mannucci, 2017).

However, deviation is not inherently beneficial. Exploratory moves may dissipate effort, lead to dead ends, or generate outcomes that are difficult to integrate into an existing knowledge base. This implies that the relationship between deviation and novelty is unlikely to be uniform across projects. Nonetheless, holding other factors constant, deviation increases the likelihood that a project moves beyond incremental refinement of the initial formulation and engages with less familiar areas of the search space. As such, deviation increases the opportunity for novelty by exposing the project to less familiar portions of the search space. We therefore hypothesize the relation between deviation and novelty as follows:

> *Hypothesis 1: Greater deviation from an initial problem formulation is associated with a higher likelihood of novel outcomes*

Coherence as a boundary condition for the value of deviation

Whether deviation translates into novelty depends critically on the structure of the search space from which exploration begins. We treat coherence primarily as anchoring in the idea generator's prior knowledge base (rather than internal consistency of the plan itself). This anchoring provides the cognitive scaffolding for interpreting and integrating exploratory outcomes, thus conditioning the value of the deviation.

Highly coherent formulations provide a dense cognitive foundation that helps actors interpret unexpected findings and integrate new insights. When deviation occurs within such a structured space, exploratory moves remain intelligible and can be recombined with existing knowledge more readily. In contrast, weakly anchored formulations provide limited guidance for evaluating and integrating exploratory outcomes. In such cases, deviation may expand the search space without providing the cognitive scaffolding needed to convert exploration into novelty.

This argument reconciles two strands of prior work that are often treated separately. Research emphasizing coherence and cumulative expertise highlights the benefits of structured search (Kaplan & Vakili, 2015), while prior work emphasizing exploration underscores the creative potential of deviation (Fleming, 2001). We argue that deviation is most valuable when these forces operate jointly: deviation generates novelty when it unfolds within a coherent cognitive map that enables integration. We thus hypothesize the moderating role of coherence as follows:

> *Hypothesis 2: The positive association between deviation from an initial problem formulation and novelty is stronger when the formulation is more coherent*

Team-based knowledge recombination and the scope of search

Exploratory search in innovation is often a collective process, shaped by changes in team composition over time. New team members introduce additional knowledge that expands the search space and influences how deviation unfolds (Teodoridis et al., 2019; Mindruta et al., 2024). We focus on the scope of recombination associated with new collaborators, distinguishing between local and distant knowledge that they introduce relative to that underpinning the initial formulation.

When new team members bring knowledge that is closely related to the original problem space, recombination remains local, reinforcing existing trajectories. In contrast, distant recombination introduces knowledge elements that lie further from the initial cognitive map, increasing the potential for novel

combinations (Fleming & Sorenson, 2004; Kneeland et al., 2020). Importantly, the value of such distant recombination depends on the coherence of the initial formulation. Coherent cognitive maps provide a structure that facilitates the absorption and integration of distant inputs, whereas weakly structured maps may struggle to incorporate them effectively.

Accordingly, we expect that within efforts that stem from coherent formulations, deviation is more likely to generate novelty when it is supported by distant rather than local recombination. Distant knowledge broadens the search space, while coherence ensures that exploratory moves remain integrable. In contrast, local recombination primarily deepens search near the initial map without increasing the likelihood of novel outcome. We thus hypothesize the moderating role of recombination scope as follows:

> *Hypothesis 3: Among projects with coherent initial problem formulations, the likelihood of novelty from deviation introduced by new team members is higher with distant than local knowledge recombination*[6]

DATA AND METHODOLOGY

Empirical Context

To examine how deviation from initial problem formulations shapes novel outcomes, we require a setting in which (i) efforts begin with clearly articulated plans, (ii) subsequent outcomes can be systematically linked back to those plans, and (iii) changes in team composition over time can be observed. We study scientific research projects conducted in an academic environment that exhibit the processes of planning, exploratory search, and team-based knowledge recombination that are central to innovation in many organizational settings. In this context, projects typically

[6] As Hypothesis 3 is conditional on a coherent starting position, we evaluate it empirically within the subset of high-coherence projects; as a robustness check, we also assess a specification including the implied three-way interaction (reported in the Appendix).

commence with a formal articulation of the problem to be addressed that outlines objectives, identifies challenges, and proposes approaches. These formulations serve as cognitive maps that guide subsequent work but do not rigidly specify research trajectories. As projects unfold, teams adapt their approaches, incorporate new insights, and often expand through the addition of new contributors. The resulting outputs (publications) reflect the realized outcome of the project, allowing us to observe the degree to which outcomes deviate from initial formulations and how team-based recombination shapes this process. Although our empirical setting is scientific research, it captures general features of innovation efforts that extend well beyond academic science. Many organizations rely on initial plans such as project proposals, development roadmaps, or investment cases, to structure exploratory work, while allowing teams discretion to adapt as projects unfold. The transparency of planning documents, outputs, and team changes in our setting allows us to observe these processes directly and to measure deviation, coherence, and recombination with unusual precision. As such, our setting provides a powerful lens for examining how planning structures exploratory search and when deviation generates novel outcomes.

Data sources and unit of analysis

Our empirical analysis links initial project formulations to subsequent outputs at the project–output level. We combine administrative records on grant-funded research proposals with publication data that identify outputs associated with each project and the researchers involved. This linkage allows us to observe for each project, the initial formulation, the degree of deviation represented by each output, and the composition of the team contributing to that output.

The core unit of analysis is a project–output pair. This structure enables us to examine variation in deviation and novelty across outputs arising from the same project, while accounting for differences in initial formulation and team composition.

Sample construction

We construct the estimation sample in several stages. We begin with the population of researchers affiliated with a leading European research university (KU Leuven) and retrieve their complete publication histories over the study period. We then identify grant proposals by these researchers that receive funding and match each project to outputs that explicitly acknowledge support from that project. This procedure yields a sample of project–output pairs in which the temporal ordering between initial formulation and subsequent output is observed.

To maintain a clear distinction between planned and realized work, we apply three restrictions. First, we retain only outputs associated with a single project, excluding cases where multiple projects jointly support the same output. This restriction avoids confounding deviation with the presence of competing or complementary plans. Second, we exclude outputs produced shortly after project start dates (1 year or less), ensuring sufficient temporal separation between initial formulation and realized outcomes. Third, we focus on projects conducted by small to medium-sized teams (consisting of 15 or fewer researchers), where individual contributions are more clearly attributable and knowledge recombination is more directly observable.

After applying these restrictions, the final estimation sample is comprised of 460 projects awarded to 976 researchers, linked to 1,341 publications authored by 4,193 individuals. Detailed information on the data sources, matching procedures, and sample construction steps is provided in Appendix 1.

Variables

Novel outcomes. Our dependent variable captures whether a project output exhibits novelty. We operationalize novelty as the introduction of new conceptual elements in an output at the time it is produced. Following Arts et al. (2025), we identify an idea as novel if it introduces terminology

that has not appeared previously in the relevant corpus and is subsequently adopted in later work.[7] This approach captures novelty as the emergence of new ideas, independent of later visibility or impact. The result is a binary indicator that equals one when an output introduces at least one such new concept and zero otherwise.[8] This operationalization aligns with prior work that emphasizes novelty as originality rather than *ex post* success and is well-suited to examining how exploratory search shapes idea generation.

Deviation from the initial problem formulation. Our central independent variable captures the extent to which an output deviates from the project's initial problem formulation. Conceptually, deviation reflects search that reorients attention away from the original cognitive map. We operationalize deviation as the semantic distance between the initial problem formulation and the realized output, based on a comparison of their textual representations. Higher values indicate greater departure from the initial formulation. This measure allows us to capture variation in the extent to which projects remain close to, or move away from, their original framing as they unfold.

Deviation. We apply natural language processing (NLP) techniques to construct a measure of the deviation between a project proposal and an associated publication, based on the abstracts.[9] Following standard text pre-processing practice, we lowercase, tokenize, remove special characters, and remove stop words to ensure comparability of the abstracts (for example, see Arts

[7] We link our publication sample to the open repository of Arts et al. (2025) using the *doi* of publications, which relies on the latest version of the Open Alex database.

[8] In robustness checks we also use a measure based on the number of times the novel terms have been reused in subsequent publications.

[9] Prior research has shown that abstracts provide a reliable representation of the main concepts, findings, and contributions contained in full texts (Hofstra et al., 2020). To ensure that the project descriptions in our data reflect scientific content rather than language used to secure funding, we follow Peng et al. (2024) and measure the use of "promotional language" in the project abstracts. As outlined in Appendix B, promotional language represents only a minor portion of each project abstract (on average, 1.4%). Our text preprocessing procedure also filters out "stop words" that lack scientific meaning, further ensuring the scientific content of the project abstracts in our data. We provide specific examples in Appendix B, Table B1.

et al. 2021).[10] Subsequently, we use pre-trained language models to create vector representations of the abstracts, utilizing the SciBERT language model (Devlin et al., 2019) and trained on the full text of publications from SemanticScholar.org (Beltagy et al., 2019).[11] We measure *deviation* as one minus the cosine similarity of the abstract of a project and its resulting publication.[12] *Deviation* can range from 0 to 1 and equals 0 when the publication is semantically equivalent to the project proposal. In our sample, the mean *deviation* is 0.29, with a standard deviation of 0.07, and a range of 0.07 to 0.54. The fields of *Economics* and *Health Sciences* show a relatively higher mean *deviation* of 0.32, suggesting a greater tendency to pivot from the project proposal than in other disciplines. In contrast, fields such as *Law and Legal Studies* and *Language and Communication Studies* exhibit a lower mean *deviation* of 0.22 and 0.23, respectively.[13] To illustrate the ability of our measure to capture the extent of deviation from the project proposal, we present two examples of project-publications pairs, together with their measure of *deviation* in Appendix 2.

Coherence of the initial problem formulation. Coherence captures the degree to which the initial problem formulation is anchored in the project leader's prior knowledge base. In line with our theoretical view of problem formulations as cognitive maps, coherence reflects how structured and internally consistent is the starting point for exploratory search. We operationalize coherence

[10] As a robustness check, we repeat our analysis of the antecedents of deviations using the ScispaCy library (Neumann et al., 2019) to generate the *deviation* metric. ScispaCy has been trained on the scientific corpus of Pubmed Central Open Access and is proficient within the biomedical domain. The pairwise correlation between the two deviation measures is positive (0.30) in our overall sample, and higher (0.35) within the biomedical and health sciences subsample. We confirm our findings using this alternative *deviation* metric.

[11] Semantic Scholar is an index over 200 million academic papers sourced from publisher partnerships, data providers, and web crawls.

[12] Based on a direct assessment of the overlap between the publication abstract and the project abstract *concepts*, we define two additional metrics of deviation: (i) the breakaway dummy, which equals 1 if no concepts in common in the publication and project abstract are found; and (ii) the inverse Jaccard index that is one minus the ratio of the size of the intersection of concepts between the publication and project abstracts to the size of their union, providing a measure of dissimilarity between the two abstracts. We define *concepts* as those after text pre-processing (i.e., net of stop words and reduced to their root form). We report these supplementary analyses in the Appendix C, Tables C7, C8 and C10.

[13] We report the distribution of *deviation* across our sample in the Appendix B, Figure B1, and show summary statistics by Fields of Research in the Appendix, Table B2.

by assessing the semantic similarity between the project's initial formulation and the project leader's prior work, using publications in the 5 years prior to the project start year. Higher values indicate that the formulation closely aligns with the leader's existing knowledge base, suggesting a coherent cognitive map, whereas lower values indicate a more weakly anchored formulation. To facilitate interpretation, coherence is standardized within funding contexts, as some funders may systematically back more exploratory projects, and thus allowing comparisons relative to projects supported by similar funding environments.

Knowledge recombination scope. Knowledge recombination scope captures the nature of exploratory search underlying deviation by reflecting whether projects draw on knowledge that is local or distant relative to the initial formulation. We focus on new team members who join the project after its initiation (i.e., those not part of the original grant application), as they represent a primary channel through which additional knowledge enters the project. We operationalize recombination scope by comparing the prior work of these new contributors to the project's initial formulation. When new team members' prior work (using their publications from the 5 years preceding the project start year) closely aligns with the original problem space, recombination is local; when it is less aligned, recombination is distant (Kneeland et al., 2020).[14] For ease of interpretation, we distinguish between local and distant recombination using indicator variables that identify outputs in the upper and lower tails of the similarity distribution, with intermediate cases serving as the reference category. More specifically, *Local Recombination* equals 1 if the average similarity between new team members' prior work and the project abstract falls at or above

[14] Specifically, we use compute the cosine similarity between new team members' publication abstracts dated in the 5 years preceding the project start year and the focal project abstract, and take the average similarity across all new team members.

the 75th percentile of the similarity distribution across all project–publication pairs. *Distant Recombination* equals 1 if the average similarity is at or below the 25th percentile.

Control variables

We include a set of control variables at the project leader, project, and output levels to account for alternative explanations.

At the project leader level, we control for prior creative performance using an indicator for whether the principal investigator has previously produced novel outputs (*PI_prior_novelty)*, as well as measures of prior productivity (*PI_prior_publications)* and impact (*PI_prior_cited)*. These variables capture persistent differences in creative ability and research style.

At the project level, we control for characteristics of the initial formulation and execution, including formulation length (*Project_abstract_WC*) and project duration *(Project_duration)*, which may influence both the scope for deviation and the likelihood of novelty. We also account for whether the project is led by a single individual or multiple leaders *(PI_solo)* and whether leaders are engaged in multiple concurrent projects (*PI_multi_fund)*.

At the output level, we control for team size (*Team_size*), the extent to which the output builds on existing work *(PriorArt_Embeddedness)*, and abstract length (*Publication_abstract_WC*).

All models include year fixed effects and domain fixed effects to account for temporal trends and systematic differences across contexts. Detailed definitions and construction procedures for all control variables are provided in Appendix 3.

Correcting for Endogeneity

To examine the relationship between deviation and novelty, we employ the control function approach (CFA) outlined by Wooldridge (2015). This two-step procedure allows us to address

omitted variable bias that may render deviation endogenous to novelty. Our intuition for using the control function approach is that unobserved research difficulties may cause OLS estimates of deviation on novelty to be downward-biased for two reasons. First, research difficulties are likely to reduce the likelihood of producing novel results. Second, these difficulties may induce scientists to deviate more from the initial plan as they adjust experimental paths or search for alternative solutions. Hence, we expect projects that encounter greater difficulty to both deviate more and achieve lower novelty, all else equal.

To implement the CFA approach, we estimate the determinants of *deviation* in the first stage equation including an "instrumental proxy" (Blundell et al., 2013) that captures exogenous sources of variation in a project's potential to change direction. We use two instruments that reflect two distinct and complementary influences on the potential for deviation that are plausibly unrelated to unobserved research difficulties affecting novelty. The first proxy captures linguistic characteristics of the project proposal that reflect the project's initial openness to change (*project openness*). The second proxy captures the latitude provided by the funding source to adjust the research direction during project execution, based on the total number of projects exhibiting substantial deviation by field of research (*funder latitude*). Funding characteristics such as monitoring practices or proposal requirements can influence the scope for deviation (Azoulay et al., 2011).

A key assumption in our econometric framework is that these instrumental proxies do not systematically correlate with unobserved research difficulties that influence novelty. While a direct econometric test of exogeneity is not feasible, our first-stage results show that the instruments strongly explain variation in our endogenous explanatory variable, deviation. Their joint explanatory power exceeds commonly used thresholds for weak identification, and we report the associated diagnostic statistics in the Appendix. The overidentification test further indicates that

the instruments are not jointly correlated with the residual variation in novelty once deviation is accounted for.

A second potential source of omitted variable bias is unobserved quality differences among researchers. Collaboration patterns in autonomous research teams often reflect assortative matching by scientific quality (Mindruta et al., 2024; Rivera et al., 2010) , which may influence both the tendency to deviate and the ability to produce novel outcomes. We mitigate this concern by controlling the project holder's prior novelty performance as well as prior publication and citation rates. In addition, we include funding source fixed effects in all regressions to account for systematic differences in funder policies regarding adherence to original research plans.

Specifically, our instrumental variables are defined as follows:

Project Curiosity. Drawing on the literature that has developed linguistic indicators to analyze scientific writing (Millar et al., 2023; Van Den Besselaar et al., 2022), we measure the degree of curiosity within the research proposal, as this may affect the project's potential to deviate from its initial plan. Our reasoning is that some proposals, motivated to advance fundamental knowledge, are open-ended in nature and have a broad scientific scope, thus allowing for exploratory approaches and a wider range of potential outcomes. In contrast, other proposals may be more narrowly focused, with a limited number of planned experiments that restrict the ability to pursue alternative research directions. To capture project-level heterogeneity in the potential to deviate we measure expressions linked to curiosity in the project proposal (Boyd et al., 2022), which may signal higher propensity to reorient the research direction, we use the software package Linguistic Inquiry and Word Count (LIWC22) to analyze the project abstract text. Our curiosity proxy

corresponds to the percentage of words in the project abstract that match one of 76 curated word entries reflecting the authors' search for, or interest in, new knowledge.[15]

Funder Latitude. The second instrumental proxy captures variation in the institutional environment that shapes how much deviation is exercised. We measure funder latitude as the number of high-deviation projects that a given funder has supported within the same field of research. This measure reflects differences in monitoring practices, tolerance for exploratory approaches, and flexibility in evaluating project outputs. Funders with high latitude are more likely to offer conditions that allow researchers to deviate from initial plans, while funders with lower latitude may implicitly constrain the scope for deviation. As this proxy varies at the funder-field level, it is not expected to correlate with project-specific research difficulties that can influence novelty outcomes.

We report on the main descriptive statistics in Tables 1 and 2 and highlight a few key observations here. The summary statistics at the publication-project level in Table 1 show that, on average, publications in our estimation sample list 1.2 project holders and 3.9 researchers joining as co-authors after the project award. The average project in our study is awarded to 2.5 scientists and produces 13.2 publications (Table 2). Further characteristics of project-holders and new team members are included in Appendix A. On average, project holders have had longer academic careers than new team members at the point when they co-publish, as measured by the years since their first publication (18.1 vs. 11.3 years). They also demonstrate higher research productivity, with an average of 12.7 publications compared to 7.4 publications for new team members, and collaborate with more coauthors (17.5 vs. 8.9). Despite these differences, both groups exhibit a

[15] The LIWC dictionary contains validated psychosocial constructs that have been recently shown to influence the success rates in project applications (Van Den Besselaar et al., 2022). Specific curiosity-related terms in the list of 76 entries include 'wonder*,' 'explor*,' 'look for*,' etc., where '*' denotes a wildcard search term.

similar average yearly citation rate per publication in the three years preceding the project award (7.6 vs. 7.0). We report the pairwise correlations between the main variables used in the regression analyses in Table A3.

RESULTS

Deviation as exploratory search and novel outcomes

We begin by examining whether deviation from initial problem formulations, conceptualized as exploratory search, is associated with novel outcomes. Table 2 reports the results from models estimating the relationship between deviation and novelty. In the baseline OLS specification, deviation is not significantly associated with novelty. As discussed earlier, however, deviation may be endogenous if unobserved research difficulties simultaneously increase the likelihood of deviating from the initial formulation while reducing the probability of producing novel outcomes. To address this concern, columns (2) through (4) in Table 2 report estimates using a control function approach. Once endogeneity is accounted for, deviation is positively and significantly associated with the likelihood that a publication introduces novel concepts (see column 2). In substantive terms, a one–standard deviation increase in deviation is associated with a meaningfully higher probability that a project produces a novel outcome, comparable in magnitude to the effect of established predictors such as the project leader's prior novelty. This result supports Hypothesis 1 and suggests that deviation, as exploratory search beyond the initial formulation, increases the opportunity for novelty.

Among the control variables, prior novelty by the project leader is positively associated with publication novelty, consistent with persistence in creative performance. In contrast, a higher prior publication rate is negatively associated with novelty, suggesting that productivity and novelty capture distinct dimensions of research behavior. Together, these findings establish

deviation as search behavior with implications for novel outcomes, laying the foundation for examining when this relationship is strengthened or weakened.

Coherence of the initial problem formulation as a boundary condition

We next test whether the coherence of the initial problem formulation conditions the relationship between deviation and novelty. Hypothesis 2 predicts that deviation is more strongly associated with novelty when the initial formulation is more coherent, i.e., more strongly anchored in the project leader's expertise. Columns (3) and (4) of Table 2 introduce coherence and its interaction with deviation. Coherence itself exhibits a small positive association with novelty, indicating that projects grounded in prior knowledge are more likely to generate novel outcomes on average. More importantly, the interaction between deviation and coherence is positive and statistically significant (see column 4). This pattern indicates that the positive association between deviation and novelty strengthens as the coherence of the initial formulation increases, consistent with Hypothesis 2.

Substantively, these results suggest that deviation does not pay off uniformly across projects. When initial problem formulations are weakly anchored, exploratory search yields limited novelty. In contrast, when projects begin from coherent cognitive maps, deviation is more likely to translate into novel outcomes.[16]

Knowledge recombination scope and the value of deviation

Finally, we examine how the scope of knowledge recombination introduced by new team members shapes the relationship between deviation and novelty. Hypothesis 3 predicts that, within projects

[16] To further assess whether the moderating role of coherence varies across different regions of the coherence distribution, we conducted additional analyses distinguishing between low- and high-coherence projects. The interaction effect is substantively large: at higher levels of coherence, the marginal effect of deviation on novelty is several times larger than at low coherence, indicating that coherence materially conditions the returns to exploratory search. These analyses, reported in Appendix 4, are consistent with Hypothesis 2 and support the interpretation of coherence as a boundary condition rather than indicating qualitatively distinct regimes.

with coherent initial problem formulations, deviation is more strongly associated with novelty when it involves distant rather than local knowledge recombination. Table 3 reports the results of models estimated on the subset of 656 projects within the top-quartile of coherence. Deviation remains positively and significantly associated with novelty across all specifications. Neither local nor distant recombination exhibits a significant main effect on novelty. However, the interaction between deviation and distant recombination is positive and statistically significant, whereas the interaction between deviation and local recombination is not. Among coherent projects, the magnitude of the interaction between deviation and distant recombination implies that deviation accompanied by distant knowledge inputs yields a substantially greater increase in novelty than deviation supported by local recombination.

These findings indicate that, among projects that begin from coherent formulations, deviation becomes more productive when it is supported by the integration of distant knowledge through new collaborators. In contrast, incorporating collaborators whose prior work closely overlaps with the initial problem formulation does not meaningfully alter the relationship between deviation and novelty. This pattern underscores that it is not recombination per se, but the scope of recombination underlying exploratory search, that conditions the value of deviation.

DISCUSSION AND CONCLUSION

Innovation projects are often initiated through plans that articulate intended goals, problems, and approaches. At the same time, novelty frequently emerges when actors adapt, reinterpret, and move beyond these initial formulations. Our study examines when and how deviations from initial problem formulations translate into novel outcomes, and how this process depends on the structure of the initial formulation and the scope of subsequent knowledge recombination.

Our first set of findings shows that deviation from an initial problem formulation is positively associated with novelty once endogeneity is addressed. This result supports the view of deviation as exploratory search that expands the opportunity for novel outcomes, rather than as mere noise or failure to execute plans. Importantly, the absence of a significant association in naïve OLS specifications highlights that deviation is not inherently beneficial; without accounting for underlying research difficulties, its value may be obscured. This finding suggests treating deviation as a substantive search behavior rather than as an outcome-neutral adjustment.

Our second set of findings demonstrates that the payoff to deviation depends on the coherence of the initial problem formulation. Deviation is more strongly associated with novelty when projects originate from formulations that are well-anchored in the project leader's prior expertise. This result reconciles seemingly competing views in the literature: while exploration and flexibility are essential for innovation, they are most effective when embedded within structured cognitive maps that enable interpretation and integration of new insights.

By conceptualizing initial plans as cognitive maps our findings challenge the notion that structure and exploration are opposing forces. Instead, structure appears to enhance the value of exploratory search by facilitating cumulative recombination. This insight advances our understanding of problem formulation and search by identifying coherence as a boundary condition for the effectiveness of deviation.

Our third set of findings highlights the role of team-based knowledge recombination in shaping the value of deviation. Among projects with coherent initial formulations, deviation generates greater novelty when it is supported by distant rather than local recombination through new team members. This result suggests that coherent cognitive maps are better able to absorb and integrate distant

knowledge inputs, allowing exploratory search to expand the solution space without losing direction.

Taken together, our study makes three contributions. First, it reframes deviation from plans as exploratory search whose value depends on the structure of the initial problem formulation. Second, it identifies coherence as a critical boundary condition that determines whether deviation yields novelty. Third, it shows how team-based knowledge recombination shapes the effectiveness of exploratory search, highlighting the complementary roles of coherence and distant recombination.

For practitioners and organizations, these findings suggest that investing in well-structured initial problem formulations need not constrain innovation. On the contrary, such formulations provide the cognitive scaffolding that allows teams to benefit from deviation and to integrate distant knowledge productively. Allowing deviation without coherence, however, may lead to exploration with limited returns.

Limitations and future research

Our study has several limitations that point to opportunities for future research. Although we address endogeneity using a control function approach, unobserved heterogeneity may still influence the relationship between deviation and novelty. Future work could exploit exogenous shocks to project trajectories or changes in planning regimes to strengthen causal inference. In addition, our focus on problem formulations and team recombination abstracts from other causes of deviation, such as serendipitous discoveries or competitive dynamics. Qualitative and mixed-method studies could complement our approach by tracing how teams interpret and revise cognitive maps over time. Finally, while our setting provides a rich environment for linking proposals,

teams, and resulting outputs, enhancing the generalizability of our findings requires consideration of other organizational and institutional contexts. For example, in a corporate R&D setting, researchers may lack the autonomy to deviate extensively from initial problem formulations. Future research could more finely distinguish between opportunity-driven and necessity-driven deviation, or examine how project leaders strategically recruit new collaborators with particular skills or expertise.

Table 1: Summary statistics (publication-project pairs, N=1,144)

Variable	Definition	Mean	SD	Min	Max
Novelty (0/1)	1 if the publication introduces at least one new scientific term	0.07	0.26	0.00	1.00
Deviation	Inverse cosine similarity between publication abstract and project abstract	0.29	0.07	0.07	0.54
Coherence	Similarity between idea generator's prior publications (past 5 years) and project abstract	0.78	0.09	0.36	0.96
Local_Recombination (0/1)	1 if avg. similarity between new team members' prior work and project abstract ≥ 75th percentile	0.25	0.43	0.00	1.00
Distant_Recombination (0/1)	1 if avg. similarity between new team members' prior work and project abstract ≤ 25th percentile	0.23	0.42	0.00	1.00
PI_prior_novelty (0/1)	1 if idea generator introduced at least one new concept in prior work (past 5 years)	0.08	0.27	0.00	1.00
PI_prior_cited	Coauthor-weighted yearly average citations per publication (past 5 years)	0.81	1.33	0.00	9.58
PI_prior_publications	Coauthor-weighted yearly average publications (past 5 years)	2.55	2.15	0.00	20.70
PI_multi_fund (0/1)	1 if idea generator has other active projects not acknowledged in focal publication	0.40	0.49	0.00	1.00
PI_solo (0/1)	1 if there is a single idea generator without additional project holders	0.64	0.48	0.00	1.00
Team_size	Number of authors on the publication	5.21	2.90	1.00	15.00
PriorArt_Embeddedness	Number of references in publication relative to the average number of references in same field	0.98	0.57	0.00	3.71
Publication_abstract_WC	Publication abstract word count (in hundreds of words)	0.94	0.34	0.16	2.79
Project_abstract_WC	Project abstract word count (in hundreds of words)	3.65	2.78	0.54	15.40
Project_duration	Years between project start and end date	4.27	1.13	0.92	7.48
Curiosity	Percent of curiosity-related words in the project abstract	0.74	0.92	0.00	9.64
Funder_Latitude	Number of high-deviation projects funded in same field by the same funder	10.90	9.16	0.00	31.00

Table 2: Deviation from Initial Problem Formulations, Coherence, and Novel Outcomes

	Novelty (0/1)			
	OLS	Control Function		
	(1)	(2)	(3)	(4)
Deviation	0.005	0.148***	0.138***	0.126***
	[0.008]	[0.050]	[0.047]	[0.047]
Coherence			0.028*	0.019
			[0.016]	[0.016]
Deviation **x** Coherence				0.025**
				[0.012]
PI_prior_novelty (0/1)	0.103**	0.122**	0.120**	0.115**
	[0.047]	[0.057]	[0.056]	[0.055]
PI_prior_cited (ln)	0.001	0.008	0.008	0.007
	[0.009]	[0.012]	[0.011]	[0.011]
PI_prior_pubs (ln)	-0.020*	-0.032*	-0.031*	-0.031**
	[0.012]	[0.016]	[0.016]	[0.015]
PI_multi_fund (0/1)	0.007	0.019	0.017	0.017
	[0.017]	[0.019]	[0.019]	[0.018]
Team_size (ln)	0.002	-0.013	-0.008	-0.008
	[0.016]	[0.020]	[0.019]	[0.019]
PriorArt_Embededdness	-0.027**	0.014	0.006	0.005
	[0.013]	[0.020]	[0.018]	[0.017]
Publication_abstract_WC	0.110***	0.119***	0.122***	0.124***
	[0.037]	[0.041]	[0.042]	[0.042]
Project_abstract_WC	-0.005	0.000	0.003	0.002
	[0.004]	[0.006]	[0.006]	[0.006]
Project_duration	0.011*	0.008	0.007	0.007
	[0.006]	[0.010]	[0.010]	[0.010]
PI_solo (0/1)	0.015	0.007	0.003	0.007
	[0.026]	[0.033]	[0.031]	[0.031]
Constant	0.146	0.173	0.176*	0.178*
	[0.153]	[0.108]	[0.096]	[0.099]
CF(Deviation)		-0.146***	-0.138***	-0.126***
		[0.050]	[0.047]	[0.047]
CF(Deviation) x Coherence				-0.032**
				[0.014]
Observations	1144	1144	1144	1144
RMSE	0.251	0.277	0.272	0.268

Notes: Columns (2)-(4) Estimates from a two-step control function approach; detailed first-stage results are available in Appendix E, Table E3. Deviation is measured as the inverse cosine similarity between the publication and grant abstract. We use the z-score of this measure to aid interpretability. PI denotes principal investigator(s) of the project. All models include fixed effects for Field of Research and Year of publication. Novelty equals 1 if a publication introduces at least one new word to the corpus of science, and the word is used at least twice . Robust standard errors, clustered at the grant level, are shown in brackets. Significance levels: * $p<0.10$, ** $p<0.05$, *** $p<0.01$.

Table 3: Recombination Scope, Deviation, and Novel Outcomes in High-coherence Projects

	Novelty (0/1)				
	(1)	(2)	(3)	(4)	(5)
Deviation	0.165**	0.166**	0.134**	0.154**	0.146**
	[0.067]	[0.066]	[0.060]	[0.061]	[0.062]
Distant Recombination (0/1)		0.006	0.004		
		[0.053]	[0.055]		
Deviation **x** Distant Recombination (0/1)			0.199**		
			[0.095]		
Local Recombination (0/1)				0.031	0.038
				[0.027]	[0.030]
Deviation **x** Local Recombination (0/1)					0.024
					[0.041]
PI_prior_novelty (0/1)	0.252*	0.253*	0.250*	0.249*	0.247*
	[0.146]	[0.146]	[0.134]	[0.141]	[0.140]
PI_prior_cited (ln)	0.016	0.016	0.012	0.013	0.013
	[0.015]	[0.015]	[0.014]	[0.015]	[0.015]
PI_prior_publications (ln)	-0.01	-0.011	-0.005	-0.01	-0.011
	[0.023]	[0.023]	[0.022]	[0.022]	[0.022]
PI_multi_fund (0/1)	-0.003	-0.003	-0.005	-0.003	-0.004
	[0.025]	[0.025]	[0.025]	[0.024]	[0.024]
Team_size (ln)	-0.017	-0.018	-0.017	-0.011	-0.012
	[0.026]	[0.024]	[0.023]	[0.024]	[0.024]
PriorArt_Embededdness	0.014	0.014	0.006	0.012	0.012
	[0.025]	[0.025]	[0.024]	[0.024]	[0.024]
Publication_abstract_WC	0.172***	0.173***	0.174***	0.169***	0.172***
	[0.058]	[0.058]	[0.057]	[0.057]	[0.057]
Project_abstract_WC	0.011	0.012	0.016	0.01	0.009
	[0.014]	[0.014]	[0.014]	[0.014]	[0.014]
Project_duration	0.019*	0.019*	0.019*	0.020*	0.020*
	[0.011]	[0.011]	[0.010]	[0.011]	[0.011]
PI_solo (0/1)	0.000	0.000	0.014	-0.002	-0.006
	[0.047]	[0.047]	[0.043]	[0.046]	[0.045]
Coherence	0.024	0.024	0.023	0.019	0.018
	[0.030]	[0.030]	[0.028]	[0.028]	[0.028]
Constant	0.016	0.015	-0.127	-0.001	0.003
	[0.124]	[0.129]	[0.137]	[0.119]	[0.117]
CF(Deviation)	-0.172***	-0.174***	-0.142**	-0.161***	-0.148**
	[0.066]	[0.066]	[0.060]	[0.061]	[0.061]
CF(Deviation) x Distant_Recombination			-0.193**		
			[0.088]		
CF(Deviation) x Local_Recombination					-0.045
					[0.048]
Observations	656	656	656	656	656
RMSE	0.278	0.278	0.278	0.273	0.273

Note: Columns (1)-(5) report the 2nd stage estimates from a two-step control function approach; detailed first-stage results are available in Appendix E, Table E3. Deviation is measured as the inverse cosine similarity between the publication and grant abstract. We use the z-score of this measure to aid interpretability. PI denotes the principal investigator(s) of the project. *Distant Recombination* is defined as projects in which the average similarity between new team members' prior publications in the past five years and the project abstract falls at or below the 25th percentile (conversely, at or above the 75th percentile for *Local Recombination*). All models include fixed effects for Field of Research and Year of publication. Novelty equals 1 if a publication introduces at least one new word to the corpus of science, and the word is re-used at least twice in later work. Robust standard errors, clustered at the grant level, are shown in brackets. Significance levels: * $p<0.10$, ** $p<0.05$, *** $p<0.01$.

BIBLIOGRAPHY

Arts, S., Hou, J., & Carlos Gomez, J. (2021). Natural language processing to identify the creation and impact of new technologies in patent text: Code, data, and new measures. *Research Policy*, *50*, 104144. https://doi.org/10.5281/zenodo.3515985

Arts, S., Melluso, N., & Veugelers, R. (2025). Beyond Citations: Measuring Novel Scientific Ideas and their Impact in Publication Text. *Review of Economics and Statistics*, 1–33. https://doi.org/10.1162/REST_A_01561/127741/BEYOND-CITATIONS-MEASURING-NOVEL-SCIENTIFIC-IDEAS

Azoulay, P., Graff Zivin, J. S., & Manso, G. (2011). Incentives and creativity: evidence from the academic life sciences. *RAND Journal of Economics*, *42*(3), 527–554. https://doi.org/10.1111/j.1756-2171.2011.00140.x

Baer, M., Dirks, K. T., & Nickerson, J. A. (2013). Microfoundations of strategic problem formulation. *Strategic Management Journal*, *34*(2), 197–214. https://doi.org/10.1002/SMJ.2004

Beltagy, I., Lo, K., & Cohan, A. (2019). *SCIBERT: A Pretrained Language Model for Scientific Text*. https://github.com/allenai/scibert/.

Blundell, R., Kristensen, D., & Matzkin, R. L. (2013). Control Functions and Simultaneous Equations Methods †. *American Economic Review: Papers & Proceedings*, *103*(3), 563–569. https://doi.org/10.1257/aer.103.3.563

Cromwell, J. R., & Harvey, J. F. (2025). A problem half-solved is a problem well-stated: Increasing the rate of innovation through team problem discovery. *Research Policy*, *54*(3), 105186. https://doi.org/10.1016/J.RESPOL.2025.105186

Deichmann, D., & Jensen, M. (2018). I can do that alone…or not? How idea generators juggle between the pros and cons of teamwork. *Strategic Management Journal*, *39*(2), 458–475. https://doi.org/10.1002/smj.2696

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *ArXiv:1810.04805*.

Fleming, L. (2001). Recombinant uncertainty in technological search. *Management Science*, *47*(1), 117–132. https://doi.org/10.1287/MNSC.47.1.117.10671

Fleming, L., & Sorenson, O. (2004). Science as a map in technological search. *Strategic Management Journal*, *25*, 909–928. https://doi.org/10.1002/smj.384

Gavetti, G., & Levinthal, D. (2000). Looking Forward and Looking Backward: Cognitive and Experiential Search. *Administrative Science Quarterly*, 45(1), 113–137. https://doi.org/10.2307/2666981

Haeussler, C., & Sauermann, H. (2020). Division of labor in collaborative knowledge production: The role of team size and interdisciplinarity. *Research Policy*, *49*(6). https://doi.org/10.1016/j.respol.2020.103987

Hofstra, B., Kulkarni, V. V., Galvez, S. M. N., He, B., Jurafsky, D., & McFarland, D. A. (2020). The diversity–innovation paradox in science. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(17), 9284–9291. https://doi.org/10.1073/PNAS.1915378117

Hottenrott, H., & Lawson, C. (2014). Research grants, sources of ideas and the effects on academic research. *Economics of Innovation and New Technology*, *23*(2), 109–133. https://doi.org/10.1080/10438599.2013.814425

Kaplan, S., & Vakili, K. (2015). The double-edged sword of recombination in breakthrough innovation. *Strategic Management Journal*, *36*(10), 1435–1457. https://doi.org/10.1002/SMJ.2294

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. https://doi.org/10.5465/3069433

Kelchtermans, S., & Veugelers, R. (2013). Top research productivity and its persistence: Gender as a double-edged sword. *The Review of Economics and Statistics*, *95*(1), 273–285.

Kneeland, M. K., Schilling, M. A., & Aharonson, B. S. (2020). Exploring uncharted territory: Knowledge search processes in the origination of outlier innovation. *Organization Science*, *31*(3), 535–557. https://doi.org/10.1287/ORSC.2019.1328

Lenfle, S., & Loch, C. (2010). Lost roots: How project management came to emphasize control over flexibility and novelty. *California Management Review*, *53*(1), 32–55. https://doi.org/10.1525/CMR.2010.53.1.32

March, J. G. (1991). Exploration and Exploitation in Organizational Learning. *Organization Science*, 2(1), 71–87. https://doi.org/10.1287/orsc.2.1.71

Millar, N., Mathis, B., Batalo, B., & Budgell, B. (2023). Trends in the Expression of Epistemic Stance in NIH Research Funding Applications: 1985-2020. *Applied Linguistics*, 1–18. https://doi.org/10.1093/applin/amad050

Mindruta, D., Bercovitz, J., Mares, V., & Feldman, M. (2024). Stars in Their Constellations: Great Person or Great Team? *Management Science*, *71*(3), 2170–2191. https://doi.org/10.1287/MNSC.2021.01969

Nelson, R. R. (2016). The sciences are different and the differences matter. *Research Policy*, *45*(9), 1692–1701. https://doi.org/10.1016/J.RESPOL.2015.05.014

Neumann, M., King, D., Beltagy, I., & Ammar, W. (2019). *ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing*. https://github.com/allenai/

Nickerson, J. A., & Zenger, T. R. (2004). A Knowledge-Based Theory of the Firm--The Problem-Solving Perspective. *Organization Science*, 15(6), 617–632. https://doi.org/10.1287/orsc.1040.0093

Peng, H., Qiu, H. S., Fosse, H. B., & Uzzi, B. (2024). Promotional language and the adoption of innovative ideas in science. *Proceedings of the National Academy of Sciences of the United States of America*, *121*(25). https://doi.org/10.1073/PNAS.2320066121

Perry-Smith, J. E., & Mannucci, P. V. (2017). From creativity to innovation: The social network drivers of the four phases of the idea journey. *Academy of Management Review*, *42*(1), 53–79. https://doi.org/10.5465/AMR.2014.0462

Petersen, A. M., Riccaboni, M., Stanley, H. E., & Pammolli, F. (2012). Persistence and uncertainty in the academic career. *PNAS*, *109*(14), 5213–5218. https://doi.org/10.1073/PNAS.1121429109/-/DCSUPPLEMENTAL/APPENDIX.PDF

Rivera, M. T., Soderstrom, S. B., & Uzzi, B. (2010). Dynamics of dyads in social networks: Assortative, relational, and proximity mechanisms. *Annual Review of Sociology*, *36*(Volume 36, 2010), 91–115. https://doi.org/10.1146/ANNUREV.SOC.34.040507.134743/CITE/REFWORKS

Rostamabadi, H., Reza Falsafi, S., Mahdi Rostamabadi, M., Assadpour, E., & Mahdi Jafari, S. (2021). Electrospraying as a novel process for the synthesis of particles/ nanoparticles loaded with poorly water-soluble bioactive molecules. *Advances in Colloid and Interface Science*, *290*. https://doi.org/10.1016/j.cis.2021.102384

Rotolo, D., Hopkins, M., & Grassano, N. (2023). Do funding sources complement or substitute? Examining the impact of cancer research publications. *Journal of the Association for Information Science and Technology*, *74*(1), 50–66. https://doi.org/10.1002/ASI.24726

Simonton, D. K. (1999). *Origins of Genius: Darwinian Perspectives on Creativity*. Oxford University Press.

Simonton, D. K. (2003). Scientific creativity as constrained stochastic behavior: the integration of product, person, and process perspectives. *Psychological Bulletin*, *129*(4), 475–494. https://doi.org/10.1037/0033-2909.129.4.475

Simonton, D. K. (2021). Scientific Creativity: Discovery and Invention as Combinatorial. *Frontiers in Psychology*, *12*. https://doi.org/10.3389/fpsyg.2021.721104

Sunilkumar, P., & Shaji, A. P. (2019). A Survey on Semantic Similarity. *2019 International Conference on Advances in Computing, Communication and Control (ICAC3)*. https://doi.org/10.1109/ICAC347590.2019.9036843

Thomke, S. H. (1997). The role of flexibility in the development of new products: An empirical study. *Research Policy*, *26*(1), 105–119. https://doi.org/10.1016/S0048-7333(96)00918-3

Vakili, K., & Kaplan, S. (2021). Organizing for innovation: A contingency view on innovative team configuration. *Strategic Management Journal*, *42*(6), 1159–1183. https://doi.org/10.1002/SMJ.3264

Van Den Besselaar, P., & Mom, C. (2022). The effect of writing style on success in grant applications. *Journal of Informetrics*, *16*(1), 101257. https://doi.org/10.1016/J.JOI.2022.101257

Wooldridge, J. M. (2015). Control Function Methods in Applied Econometrics. *The Journal of Human Resources*, *50*(2), 420–445. https://www.jstor.org/stable/24735991

Wu, L., Wang, D., & Evans, J. A. (2019). Large teams develop and small teams disrupt science and technology. *Nature*, *566*, 378–382. https://doi.org/10.1038/s41586-019-0941-9

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. https://doi.org/10.1177/0001839219843486

APPENDIX

- PW correlations
- 1st stage results for Tables 2
- Table 2 using IV as a validation of CF approach. Include the diagnostics of IV
- 1st stage results for Table 3,4
- Tables 2,3,4 using alternative measure of novelty (ln_word_reuse) number of reuse of new word, instead of novelty as a dummy
- Table 4 for low coherence projects
- TO BE CHECKED WHAT ADDITIONAL ANALYSES / APPENDIX FROM THE PREVIOUS VERSION ARE STILL USEFUL/PERTINENT:
    - Supplementary data description? Grant-level; Researcher-level; Publication supported by multiple vs single grant
    - Supplementary info on deviation: examples , promotional language, etc
    - Supp analyses , funding source heterogeneity ?
    - Supp analyses: grant fixed effect?
    - Supp analyses: citations?

Appendix 1. Details on the Sample Construction

As the basis for our analysis, we start from funded research projects conducted by scientists at KU Leuven in Belgium.[17] We combine several sources into a unique dataset containing grant-publication pairs to identify idea generators (the principal investigators who were awarded project funding) and identify additional researchers included as coauthors in the publications (coauthors who joined the project post-grant award).

This setting has several desirable features for our analysis. First, project-based funding is a cornerstone of research budget allocation across all academic disciplines at KUL. Scientists routinely submit project proposals to obtain necessary funding for equipment and to hire technical staff, Ph.D. students, and postdoctoral researchers. Second, all funded projects are centrally reported to the Flemish government by the university-wide research coordination office, thus allowing us to reliably identify and link researchers, funding records, and publication data.

We identify the organizational affiliation of each researcher at the time of publication and project award, as well as past collaborations among project holder(s) and coauthors joining after the project award. Our data originates from three sources. From university records, we obtained personnel data across all scientific fields represented at the university (social sciences and humanities, natural sciences and engineering, and biomedical and life sciences) in the period 2000-2018. Second, we obtained data on publications and project funding from the Dimensions database of Digital Science. Funding information in Dimensions is provided directly by the funding

[17] KU Leuven is a research-intensive university with recruitment, promotion, and research policies conforming to international practices in the pursuit of research excellence and scientific impact. KUL is a founding member of the League of European Research Universities (LERU) and a member of international university associations including the Coimbra Group and UNA Europa, among others. The university ranked 42nd in the Times Higher Education (THE) ranking in 2022, is the highest-ranked institution in Belgium according to THE, and has been consistently ranked among the top 50 universities globally in recent years.

organizations, especially when funders are partners of Digital Science.[18] Our third source is the official public funding records in Flanders, Belgium (the "FRIS data"), which are available from 2008 onwards. The FRIS database includes all funded research in our setting, where funding is distributed through competitive calls from various sources, including regional and federal sources in Belgium (e.g., the Research Foundation Flanders and the Belgian Science Policy Office), the European Union (e.g., the ERC), and other international sources such as funding agencies outside Belgium.[19]

We began by curating a list of 6,907 KUL researchers and retrieving publication records for each researcher since 1980. For each publication, we obtain metadata including the title, abstract, coauthors and their affiliation(s), publication date, disciplines (fields of research)[20], references, acknowledged funding information, and citation metrics as of December 2022. We obtain similar publication data for all coauthors of our initial set of KUL researchers.

From the publication corpus of 202,354 publications, 36,994 (18%) cite a supporting grant. For each project, we obtain information such as the project title, abstract (i.e., a summary of the project proposal), funder details, project start date and end date, researchers who were awarded the project, and their affiliation(s) from the Dimensions database. From Dimensions, we also capture all projects awarded to our curated list of researchers and their co-awardees. We match these records with data from the FRIS database based on project title, funder details, and researcher

[18] Data on projects reported in the Dimensions database are provided directly by partner funding bodies and complemented by public sources, covering more than 600 funders worldwide including governments, nonprofit organizations, and international agencies. Dimensions further complements these data by processing project codes mentioned in the acknowledgement section of publications, and then matched to the Dimensions project database, in line with prior work in the literature (Grassano et al., 2017).

[19] A limitation of the FRIS data is that linkages to funded project's output are incomplete. We use the FRIS data primarily to complement information on projects retrieved from Dimensions (e.g., to validate the original team of project holders and fill in missing values on project abstracts, start and end dates, etc.).

[20] Fields of Research provided by Dimensions follow the Australian and New Zealand Standard Research Classification (ANZSRC 2020), composed of 22 disciplines.

names. Of the grant-supported publications, 9,183 publications acknowledge financial support awarded to (at least one) KUL researcher(s) in the years 2000-2018.[21] After initial data cleaning, we are able to collect full information for 1,107 projects linked to 6,196 publications, representing 67% of the grant-supported publications in our study period.

We apply three restrictions to this sample of project-publication pairs. First, to ensure a clear point of comparison of the deviation of a given publication from the initial project proposal, we include only publications supported by a single project.[22] Previous studies indicate that the coexistence of multiple funding sources creates synergies and tensions between different project objectives that may alter the research process, and that are difficult to disentangle in the resulting publication output (Rotolo et al., 2023). Moreover, multiple funding sources can influence both the sources of new ideas and the scientific impact of publications (Hottenrott and Lawson, 2014), thus may confound the effect of changes in team composition that we aim to uncover.

Second, as our focus lies on research that is actually supported by grants, we exclude publications with a date of publication of less than a year from the starting date of the project. Instances where the project proposal describes research that has already largely been performed may blur the temporal separation between the research plan outlined in the project proposal and its resulting publications. In turn, this could obscure the "true" deviation between the proposed research and its publication output, potentially leading to an artificially low deviation rather than describing new research.[23]

[21] The remaining publication-project pairs are deemed out of scope since they have been awarded in a date before our observation period (2000-2018) or are linked to a group of scientists that is entirely external to KU Leuven.
[22] See Appendix A for supplementary data description. In supplementary analyses, we extend our sample and include publications supported by multiple projects (Appendix C, Table C3). As expected, we find a positive and statistically significant coefficient for the number of supporting projects mentioned by a given publication. We also find supporting evidence on the positive correlation between the presence of new team members and deviation from project proposals.
[23] We further substantiate our chosen threshold in the Appendix A. In a supplementary analysis, we modify the one-year threshold and only exclude publications appearing within two months from the project start date. Our findings are largely insensitive to this adjustment (Appendix E, Table E6).

Third, to maintain empirical and theoretical tractability, we restrict our analysis to teams of fifteen or fewer scientists.[24] This restriction fits with our knowledge recombination logic that applies to small to medium-sized teams where individual contributions are more significant (Wu et al., 2019). Our choice of this team size threshold is consistent with prior research, particularly in studies investigating knowledge recombination processes within scientific teams (e.g., Haeussler and Sauermann, 2020).

After applying these restrictions, our estimation sample includes 460 projects awarded to 976 researchers, linked to 1,341 publications authored by 4,193 researchers.

[24] Seventy-six observations, or 3.5% of the remaining publications, exceed our team size threshold of fifteen coauthors.

Appendix 2. Details on Control Variables

Researcher-level controls

The insights and discovery moments that drive scientific progress are often associated with scientists' creativity, intuition and talent. More talented scientists are better able to generate novel ideas and to bring these ideas to fruition (Simonton, 1999, 2003). Prior work also shows that research performance is highly persistent over time, with a small group of scientists repeatedly contributing a disproportionate share of influential work (Kelchtermans and Veugelers, 2013; Petersen et al., 2012). To account for these differences, we include several controls that capture the prior performance of the idea generator, that is, the project holder responsible for the initial problem formulation.

First, we include an indicator for prior novelty that equals 1 if the idea generator introduced at least one new concept in their publications during the five years preceding the focal project (*PI_prior_novelty*). This variable captures a history of creative output and the idea generator's demonstrated ability to push the frontier. Second, we control for the idea generator's prior citation rate (*PI_prior_cited*), measured as the coauthor-weighted yearly average number of citations per publication over the same five-year window and expressed in logarithms. This measure reflects the impact of the idea generator's earlier work. Third, we include the idea generator's prior publication rate (*PI_prior_publications*), measured as the coauthor-weighted yearly average number of publications, also in logarithms, to capture productivity rather than impact. Together, these measures help separate the effect of deviation from underlying differences in the ability and track record of the idea generator.

We also add a dummy variable that equals 1 if the idea generator has other active projects that are not acknowledged in the focal publication (*PI_multi_fund*). This variable captures the possibility that attention and motivation of the PI is spread across several projects. We also include a dummy *PI_solo*, which equals 1 when there is a single idea generator without additional project holders involved in the craft of the focal project proposal. Solo responsibility may affect the way deviation is managed, so controlling for this configuration helps isolate the role of deviation itself (Deichmann & Jensen, 2018).

Project-level controls

At the project level, we control for features of the proposal and its execution that may shape both deviation and novelty. The first is the length of the project abstract *(Project_abstract_WC*), measured as the word count in hundreds of words. Longer abstracts may reflect more detailed and specific research plans, which can either constrain the scope for deviation or provide a richer basis for recombination. Second, we include project duration (*Project_duration*), measured as the number of days between the project start date and the grant end date and standardized within funder. Standardizing within funder removes mechanical differences in typical project length across funding schemes (for example, between large European Research Council grants and shorter Flemish grants) and allows us to interpret project duration as "longer or shorter than is usual for this funder." Longer projects provide more time for adjustment, exploration and recombination, which can influence both how far a project deviates from its initial formulation and the likelihood of generating novel outcomes.

Finally, funder characteristics may also differ in many other respects, such as the scope of competition for resources, and the scope of collaboration requirements for researchers (for example, the federal funding mandates collaborations across linguistic communities in Belgium).

To account for these broader differences that might otherwise confound the relationship between deviation and novelty, all regressions include the following funding source fixed effects: Science Foundation Flanders (FWO), federal funding in Belgium (Belspo), the European Research Council (ERC), other funding sources within the EU, and other funding sources outside the EU.[25]

Publication-level controls

The resulting publication reflects the collective effort of the idea generator and the broader team of researchers. We therefore include several publication-level controls that characterize the team and the way the work is positioned in the literature. Team size (*Team_size*) is measured as the number of authors listed on the publication. Larger teams may have access to a broader pool of ideas, skills and networks, which can influence both the type of recombination and the likelihood of novelty. We further control for the extent to which the publication is embedded in prior art. *PriorArt_Embeddedness* is defined as the number of references cited in the publication divided by the average number of references in publications in the same field of research. Bibliometric studies show that scientific outcomes can depend on the body of knowledge a paper builds upon and its vintage (Vieira and Gomes, 2010). Embeddedness may affect the opportunities for recombination and novelty with existing work.

Finally, we include the length of the publication abstract (*Publication_abstract_WC*), measured as the word count in hundreds of words. Longer abstracts may signal more complex or multifaceted contributions and thus a higher likelihood of containing novel elements. All models also include publication year fixed effects to capture temporal trends in research, as well as research domains fixed effects based on Fields of Research (FoR) codes. FoR is sourced from the

[25] While our data also distinguishes funding at the more fine-grained level program-level (within a funding agency), the distribution of project funding sources closely aligns with the funding agency level in our estimation sample. For example, within the publications supported by projects awarded by the FWO, 90% (299 out of 329) correspond to the 'FWO – Research projects' program.

Dimensions database where each publication is assigned at least one of 22 FoR codes, following the ANZRC 2020 classification system. We add an additional category for cases where multiple FoR codes are assigned. We follow the widely accepted view that the norms and organization of scientific research differ by domain (Nelson, 2016): some domains are more prone to open-ended research, while experimental fields expect project proposals to present a detailed plan.

Appendix 3: Examples of Deviation from Initial Problem Formulations

In Example 1 (see Table X), we show a publication-project pair with a relatively low *deviation* (0.07), i.e., the publication belongs to the bottom percentile of the *deviation* distribution and closely aligns with the original project plan. The stated goal of the project is to gain fundamental knowledge into the process and formulation parameters that influence electrospraying to improve drug delivery methods.[26] Both the project and publication emphasize the challenges of low solubility in drugs, the use of amorphous solid dispersions, and the exploration of electrospraying for particle preparation. The publication identifies key parameters of this process, such as flow rate, electric potential difference, distance between the nozzle and collector, solvent type, and polymer concentration, that are critical for the design of electrospraying methods for amorphous solid dispersions. A casual reading suggests that the results reported in the publication largely mirror the plans described in the project proposal.

In Example 2 (see Table X), we present a publication-project pair characterized by a relatively high *deviation* (0.35) —above the 75th percentile— indicating a greater disparity between the project proposal and the resulting publication. The project application describes research aimed at developing innovative materials using rod-shaped cellulose and platelet-like starch particles, focusing on controlling their arrangement through chemical modifications and exploring interactions with substances like metal nanoparticles. In contrast, the publication abstract delves into specific interactions between a particular animal protein, bovine serum albumin (BSA), and various cellulose particles. While both abstracts pertain to cellulose modification and the study of interactions, the project abstract focuses on creating new materials and understanding self-

[26] Electrospraying is an electro-hydrodynamic process that uses an electrostatic force to spin or spray a polymer solution into ultrafine droplets (Rostamabadi et al., 2021).

assembly, whereas the publication concentrates on BSA interactions, detailing the behavior of individual amino acids on cellulose surfaces. The project abstract appears to have a broader scope related to material design, while the publication abstract focuses on a particular set of experiments and findings. Importantly, the project proposal did not anticipate studying BSA, let alone any other animal proteins, indicating a notable deviation from the original research plan.

Appendix 4. Functional-form and Robustness Checks for the Moderating Role of Coherence

In the main analyses, we theorize and test the coherence of the initial problem formulation as a continuous moderator of the relationship between deviation and novelty (Hypothesis 2). To assess whether this moderating relationship reflects qualitatively distinct regimes at different levels of coherence, or instead operates as a boundary condition that attenuates or strengthens the value of deviation, we conducted additional analyses separating projects into low- and high-coherence regions of the distribution. These analyses are not intended as tests of separate theoretical mechanisms, but rather as robustness checks examining the functional form of the interaction between deviation and coherence. As reported in Table A3, the results are consistent with the main findings: deviation is less productive when coherence is low, while the positive association between deviation and novelty strengthens as coherence increases. Importantly, these results do not suggest a sharp discontinuity between regimes, supporting our decision to model coherence as a continuous boundary condition in the main theory and analyses.

Table A.1: Deviation from Initial Problem Formulations and Novel Outcomes, Low- versus High-Coherence Projects

	Novelty (0/1)			
	(1)	(2)	(3)	(4)
Deviation	0.150***	0.160***	0.142***	0.130***
	[0.052]	[0.052]	[0.048]	[0.048]
Low Coherence (0/1)	-0.049*	-0.029		
	[0.027]	[0.026]		
Deviation **x** Low Coherence (0/1)		-0.076**		
		[0.037]		
High Coherence (0/1)			0.025	0.032
			[0.025]	[0.027]
Deviation **x** High Coherence (0/1)				0.035
				[0.041]
PI_prior_novelty (0/1)	0.122**	0.123**	0.123**	0.122**
	[0.057]	[0.053]	[0.057]	[0.056]
PI_prior_cited (ln)	0.011	0.011	0.008	0.009
	[0.012]	[0.012]	[0.012]	[0.012]
PI_prior_publications (ln)	-0.025	-0.026*	-0.029*	-0.030*
	[0.016]	[0.016]	[0.016]	[0.016]
PI_multi_fund (0/1)	0.019	0.019	0.019	0.018
	[0.020]	[0.019]	[0.019]	[0.019]
Team_size (ln)	-0.008	-0.008	-0.011	-0.011

	[0.019]	[0.019]	[0.020]	[0.019]
PriorArt_Embededdness	0.014	0.01	0.011	0.011
	[0.020]	[0.020]	[0.019]	[0.019]
Publication_abstract_WC	0.117***	0.116***	0.119***	0.121***
	[0.041]	[0.041]	[0.041]	[0.041]
Project_abstract_WC	0.002	0.001	0.001	0.000
	[0.006]	[0.005]	[0.006]	[0.006]
Project_duration	0.008	0.008	0.007	0.007
	[0.010]	[0.010]	[0.009]	[0.009]
PI_solo (0/1)	0.000	-0.002	0.007	0.009
	[0.033]	[0.031]	[0.032]	[0.031]
Constant	0.162	0.139	0.161	0.158
	[0.104]	[0.105]	[0.109]	[0.111]
CF(Deviation)	-0.148***	-0.160***	-0.140***	-0.123**
	[0.052]	[0.051]	[0.048]	[0.048]
CF(Deviation) x Low Coherence		0.089**		
		[0.042]		
CF(Deviation) x High Coherence				-0.058
				[0.046]
Observations	1144	1144	1144	1144
RMSE	0.278	0.275	0.274	0.274

Notes: Columns (1)-(4) report the 2nd stage estimates from a two-step control function approach; detailed first-stage results are available in Appendix E, Table E3. Deviation is measured as the inverse cosine similarity between the publication and grant abstract. We use the z-score of this measure to aid interpretability. PI denotes the principal investigator(s) of the project. *Low Coherence* is defined as projects in which the similarity between the PI's prior work and the focal project falls at or below the 25th percentile (conversely, at or above the 75th percentile for *High Coherence*). All models include fixed effects for Field of Research and Year of publication. Novelty equals 1 if a publication introduces at least one new word to the corpus of science, and the word is re-used at least twice in later work [ADD REF]. Robust standard errors, clustered at the grant level, are shown in brackets. Significance levels: * p<0.10, ** p<0.05, *** p<0.01.