Working Paper 24-064

Greenlighting Innovative Projects: How Evaluation Structure Shapes the Perceived Feasibility of Early-Stage Ideas

Jacqueline N. Lane
Simon Friis
Tianxi Cai
Michael Menietti
Griffin Weber
Eva C. Guinan

Harvard Business School

Greenlighting Innovative Projects: How Evaluation Structure Shapes the Perceived Feasibility of Early-Stage Ideas

Jacqueline N. Lane
Harvard Business School
D^3 Institute at Harvard

Simon Friis
Harvard Business School
D^3 Institute at Harvard

Tianxi Cai
Harvard TH Chan School of Public Health

Michael Menietti
D^3 Institute at Harvard

Griffin Weber
Harvard Medical School

Eva C. Guinan
Dana-Farber Cancer Institute

Working Paper 24-064

Copyright © 2024 and 2025 by Jacqueline N. Lane, Simon Friis, Tianxi Cai, Michael Menietti, Griffin Weber and Eva C. Guinan.

Working papers are in draft form. This working paper is distributed for purposes of comment and discussion only. It may not be reproduced without permission of the copyright holder. Copies of working papers are available from the author.

Funding for this research was provided in part by Harvard Business School.

Greenlighting Innovative Projects:

How Evaluation Structure Shapes the Perceived Feasibility of Early-Stage Ideas

Jacqueline N. Lane[†*], Simon Friis[†], Tianxi Cai, Michael Menietti,

Griffin Weber and Eva C. Guinan

[†]Co-first authorship,
[*]Corresponding author

Jacqueline N. Lane
Harvard Business School &
Laboratory for Innovation Science at Harvard
Boston, MA 02163
jnlane@hbs.edu

Simon Friis
Harvard Business School &
Laboratory for Innovation Science at Harvard
Boston, MA 02163
sfriis@hbs.edu

Tianxi Cai
Harvard T.H. Chan School of Public Health
Boston, MA 02115
tcai@hsph.harvard.edu

Michael Menietti
D^3 Institute at Harvard
Boston, MA 02163
mmenietti@hbs.edu

Griffin Weber
Harvard Medical School
Boston, MA 02115
griffin_weber@hms.harvard.edu

Eva C. Guinan
Dana-Farber Cancer Institute
Boston, MA 02215
eva_guinan@dfci@harvard.edu

Greenlighting Innovative Projects:
How Evaluation Structure Shapes the Perceived Feasibility of Early-Stage Ideas

ABSTRACT

Organizations routinely use multi-criteria evaluation structures to allocate resources to innovative projects, assessing impact, novelty, and feasibility to identify initiatives worth funding. Yet even with formal feasibility assessments, many innovative projects fail during implementation, often after significant investment. Could multi-criteria evaluation—by requiring evaluators to weigh multiple dimensions simultaneously—systematically bias the very feasibility judgments it is designed to inform? We investigate this question through a field experiment at a leading research university, in which experts were randomly assigned to evaluate scientific grant proposals using either a multi-criteria (impact, novelty, and feasibility) or a single-criterion (feasibility only) rubric. Compared to multi-criteria evaluators, feasibility-only evaluators assigned systematically lower feasibility scores to innovative projects. This pattern reflects cross-dimensional spillover: when evaluators assessed all three dimensions, their impact and novelty judgments substantially influenced their feasibility scores—a spillover effect largely absent in single-criterion evaluation. These patterns suggest that multi-criteria evaluation may lead organizations to greenlight initiatives despite missed implementation barriers, as enthusiasm for novelty and high impact elevates perceived feasibility. Evaluation structures, therefore, are not neutral administrative tools but mechanisms that shape organizational attention and judgment. For organizations seeking to improve early-stage project selection, single-criterion feasibility evaluations may sharpen the detection of implementation risks compared to multi-criteria assessment, while multi-criteria approaches may better capture interdependencies across project dimensions.

Keywords: innovation evaluation, evaluation criteria, feasibility assessment, attention allocation, cognitive mechanisms, field experiment

INTRODUCTION

Despite rigorous evaluation processes, many innovative projects fail—not because their ideas lack merit, but because subtle implementation barriers derail execution (Briel et al., 2016; Fogel, 2018; Khanna et al., 2016). These feasibility issues often could have been detected during evaluation, before organizations committed substantial resources. Why do evaluators overlook such challenges? We propose that the answer lies in how organizations structure evaluation. Across domains, organizations routinely rely on multi-criteria evaluation structures that simultaneously assess impact, novelty, and feasibility (Boudreau et al., 2016; Criscuolo et al., 2017; Knudsen & Levinthal, 2007). While this comprehensive approach aims to capture the multidimensional nature of innovation, it may inadvertently dilute evaluators' focus: when assessing multiple dimensions simultaneously, divided attention and cross-dimensional spillovers can elevate feasibility judgments for projects perceived as especially novel or high-impact. If so, standard practices may not merely *reveal* which innovations fail but *contribute* to those failures by systematically biasing feasibility assessments (Fleming, 2001; Khanna et al., 2016).

This paper investigates whether the evaluation structure affects feasibility assessment. Specifically, we compare a multi-criteria (impact, novelty, feasibility) evaluation structure with a single-criterion (feasibility-only) structure for identifying potential feasibility issues in early-stage projects. We define impact as the extent to which a project can create significant value and influence key stakeholders across multiple domains, including markets, industries, and scientific fields (Kuhn, 1977; Vakili & Kaplan, 2015), novelty as the extent to which a project is original and departs from the existing knowledge frontier (Boudreau et al., 2016; Carnabuci & Operti, 2013; Kuhn, 1977; Weitzman, 1998), and feasibility as the extent to which the proposed project can be implemented within specific constraints (Franzoni & Stephan, 2023; Lane et al., 2022).

Organizations frequently rely on multi-criteria evaluation structures to guide resource allocation decisions for early-stage projects (Bian et al., 2021; Boudreau et al., 2016; Criscuolo et al., 2017; Knudsen & Levinthal, 2007). For instance, the U.S. National Institutes of Health (NIH)—the primary federal agency supporting biomedical and public health research—uses a multi-criteria structure to evaluate grant proposals across dimensions corresponding to impact, novelty, and feasibility (*NIH*, 2024). In 2023 alone, this approach guided the allocation of $34.9 billion in funding across nearly 59,000 grants (*NIH,* 2024), significantly shaping the trajectory of entire scientific fields. Yet a persistent puzzle in innovation management is that novel projects frequently fail during implementation despite passing rigorous evaluation processes (Fleming, 2001; Khanna et al., 2016)—raising questions about whether standard evaluation practices adequately screen for feasibility constraints.

Evaluating multiple criteria simultaneously can present opportunities and challenges for evaluators' attention. Decision-makers often face time constraints on their attention (Ocasio, 1997), and the amount of attention devoted to an issue depends on the number of competing demands on decision-makers' time and the interrelationships among them (Eklund et al., 2025). While multi-criteria reviews may help evaluators perceive interdependencies between dimensions (Baker & Nelson, 2005), the divided attention may dilute focus on individual aspects of a project (Ocasio, 1997). For instance, a project's novel approach might improve feasibility by circumventing traditional implementation barriers, or its feasibility constraints might drive more novel solutions (Baker & Nelson, 2005; Prahalad & Hart, 2002). However, in domains where the feasibility of a novel solution is uncertain, a feasibility-oriented assessment may help uncover issues prior to implementation, including both fatal flaws (Camerer & Johnson, 1997) and issues that can be addressed to improve the project's ultimate success.

The rationale for single-criterion feasibility reviews becomes particularly apparent when we consider staged innovation processes (Azoulay & Li, 2020; Guler, 2007; Kerr et al., 2014). For instance, recent evidence from the pharmaceutical industry indicates that many late-stage failures could have been detected before significant resources were invested, making these failures potentially preventable (Briel et al., 2016; Fogel, 2018). Despite uncertainty in the innovation process and the potential for a project's feasibility to change over time (Dougherty, 1992), early identification of critical implementation issues can enable corrections or timely termination before substantial investment (Åstebro & Elhedhli, 2006;

Khanna et al., 2016). Despite the plausible benefits of feasibility-only reviews, it is not entirely clear from the prior literature how such assessments will shape the project attributes of the selected projects that rise to the top (Lamont, 2009). By reducing the evaluation to a single criterion, such as feasibility, evaluators adopt a more focused mindset, enabling a deeper examination of specific implementation challenges (Joseph & Wilson, 2018). While evaluators still read complete proposals, the evaluation structure shapes which aspects they attend to and how they weigh them in their judgments. Yet, directing attention away from specific dimensions may risk filtering out relevant signals about a proposal's broader implications.

While these evaluations aim to predict project outcomes, studying how evaluation structures affect assessments requires experimental manipulation of established review processes—an opportunity that rarely arises in practice. Despite abundant retrospective evidence that novel projects frequently fail due to implementation barriers (Fleming, 2001; Scherer & Harhoff, 2000; Wang et al., 2017), we know surprisingly little about whether standard evaluation practices systematically overlook these barriers—potentially causing rather than merely predicting innovation failures. To address this gap, we partnered with a U.S. medical school's translational science program to design and execute a field experiment that modified its pilot grant funding process. This setting offered an ideal context to examine how evaluation structures shape assessments of implementation challenges for three reasons: it focused on knowledge creation through scientific peer review of early-stage projects; the medical school sought to enhance its evaluation procedures; and program administrators had observed that standard multi-criteria evaluations frequently failed to identify subtle yet critical feasibility issues in translational research proposals. We randomly assigned expert evaluators to either a "multi-criteria" condition (assessing impact, novelty, and feasibility) or a "single-criterion," feasibility-only condition. Our experiment involved 97 evaluators from seven institutions assessing 47 projects, yielding 641 proposal-evaluator pairs and $500,000 in funding, offering a unique window into resource allocation under contrasting evaluation structures.

Our findings reveal that evaluation structures function as organizational attention-allocation mechanisms, exerting systematic effects on resource-allocation decisions. When evaluators focus solely on feasibility, they assign feasibility scores 6–10 percentage points lower than do multi-criteria evaluators. This effect reflects two complementary mechanisms uncovered through hybrid human-LLM analysis of evaluator comments: feasibility-only evaluation enables deeper problem discovery by concentrating cognitive resources on implementation challenges, while mitigating cross-dimensional spillovers where evaluators' enthusiasm for novel, high-impact projects elevates their feasibility scores.

These findings challenge the assumption that evaluation structures are neutral administrative procedures. Our research contributes to the innovation evaluation literature by demonstrating the micro-foundations of attention allocation: how individual-level cognitive processes translate into organizational resource allocation patterns. While prior work on strategic decision-making emphasizes how organizational architecture distributes managerial attention (Joseph & Ocasio, 2012; Joseph & Wilson, 2018; March & Simon, 1958), we show that seemingly minor procedural choices—whether to assess one criterion or three—directly alter what evaluators notice, discuss, and ultimately value. By systematically manipulating evaluation structure, we reveal that rubric design represents a consequential organizational lever for directing innovative activity.

The implications extend beyond evaluation procedures to fundamental questions about innovation portfolio management. Organizations face a strategic choice: feasibility-focused evaluation enables early detection of implementation barriers that could derail promising projects, while multi-criteria approaches capture interdependencies necessary for strategic alignment. Rather than viewing this as a binary choice, our findings suggest that evaluation structure can be strategically aligned with organizational objectives—using single-criterion assessment when rigorous feasibility screening is prioritized, multi-criteria evaluation when holistic judgment across dimensions is required, or sequential and bifurcated review systems that combine both approaches at different stages of project selection. This represents a shift from treating evaluation as a fixed organizational routine to conceptualizing it as a dynamic capability that can be optimized for different innovation challenges.

EVALUATING EARLY-STAGE INNOVATION AND MULTI-CRITERIA EVALUATION

Innovation is both the driver of long-run economic growth and a deeply uncertain, failure-prone process (Schumpeter, 1942; Rosenberg & Nelson, 1994; Fleming, 2001; Leiponen & Helfat, 2010; Bikard & Marx, 2020). Theories of creative destruction and scientific paradigms emphasize that breakthroughs arise through risky departures from existing knowledge (Boudreau et al., 2016; Kaplan & Vakili, 2015; Kuhn, 1977; Ferguson & Carnabuci, 2017; Teodoridis et al., 2019). However, early-stage innovation is uniquely uncertain: most novel projects fail, and those that succeed often demand large, risky investments (Franzoni & Stephan, 2023; Lane et al., 2021; Scherer & Harhoff, 2000). Organizations must therefore make high-stakes selection decisions long before outcomes are known, identifying which projects to pursue, defer, or abandon (Baer, 2012; Baer et al., 2013; Criscuolo et al., 2017; Azoulay & Li, 2020; Lane et al., 2022). The costs of error are asymmetric: supporting infeasible projects wastes scarce resources, while discarding feasible but unconventional projects risks missing transformative opportunities. Feasibility constraints—budget limits, overly ambitious scope, timelines, or regulatory hurdles—thus loom large, making evaluation procedures critical elements in determining which projects ultimately move forward (Guler, 2007; Sleesman et al., 2018).

Evaluating innovation is difficult because project quality is inherently multi-dimensional. Accordingly, organizations typically assess the quality of early-stage projects along three main dimensions: the novelty or originality of the approach, the potential impact or significance of outcomes, and the feasibility or viability of execution (Boudreau et al., 2016; Mount et al., 2021). Specific terminology varies across contexts. The NIH, for instance, evaluates grant proposals using significance, innovation, and approach—dimensions that map onto impact, novelty, and feasibility. These dimensions are not exhaustive, but they capture the essential concerns that repeatedly surface in organizational and scholarly accounts of innovation. Indeed, many ostensibly distinct evaluation criteria—such as scalability, generalizability, or relevance—tend to collapse into some combination of impact, novelty, and feasibility when applied in practice. This triad therefore represents a natural, widely used structure for making sense of uncertain, early-stage projects, even if the specific terminology or weighting varies across contexts.

To operationalize these dimensions, organizations rely on multi-criteria evaluation structures that instruct evaluators to assess novelty, impact, and feasibility simultaneously. Such rubrics are widely institutionalized across domains such as scientific peer review (Azoulay & Li, 2020; Boudreau et al., 2016; Criscuolo et al., 2017), venture capital investment (Bian et al., 2021), and corporate R&D management (Knudsen & Levinthal, 2007). These frameworks not only structure decision-making but also channel organizational attention, codify professional norms, and reproduce institutionalized models of quality. Yet despite their pervasiveness, we know relatively little about how design choices within these rubrics affect evaluators' judgments[1]. Even seemingly minor modifications—such as whether to collapse criteria, shift weights, or emphasize feasibility over novelty—might alter which projects organizations ultimately support.

Do Evaluation Structures Matter?

While multi-criteria evaluation is the dominant approach, organizations retain considerable discretion over how to structure the evaluation process itself. A core design choice is whether evaluators should assess multiple criteria simultaneously or focus attention on a single dimension, such as feasibility, which has important consequences for the downstream implications of project success. For instance, a single-

[1] At first glance, the multi- vs. single-criteria evaluations under consideration here appear related to various strands of research on multi-attribute decisions, but the focus is different. Multi-attribute utility (MAU) models (Huber, 1974; Keeney & Raiffa, 1976; von Winterfeldt & Edwards, 1986) concern the question of how people make trade-offs when alternatives have several attributes—for example, choosing a site location by weighing cost, distance, and amenities. The goal is to elicit stable preferences and combine attributes into an overall score. Our interest, by contrast, is not in how to aggregate fixed attributes, but in how the structure of the rubric itself affects what reviewers notice—for instance, whether asking them to judge feasibility alone versus alongside impact and novelty leads to different discoveries and judgments. Joint–separate evaluation research (Hsee, 1996; Hsee et al., 1999; Hsee & Zhang, 2004) asks whether people prefer the same option when it is evaluated side-by-side with another versus on its own—for example, how a stereo system seems better or worse when compared directly with an alternative. That work is about option presentation. By contrast, our question concerns how evaluators assess a given project when asked to apply one dimension versus multiple dimensions, regardless of how many proposals they ultimately review.

criterion structure might direct reviewers to evaluate only feasibility—considering methods, timelines, and resource requirements—without explicit attention to novelty or impact. Indeed, organizations do modify their structures over time: the NIH recently simplified its review system by reducing five criteria to three (NIH, 2024)[2], illustrating that organizations actively adjust evaluation structures to improve decision quality and strategic alignment. These changes can appear to be mere administrative procedures, yet because structures determine where evaluators allocate scarce attention, they represent a straightforward organizational lever for directing assessment toward organizational goals (Azoulay & Li, 2020; Boudreau et al., 2016; Criscuolo et al., 2017; Knudsen & Levinthal, 2007; Bian et al., 2021).

It is reasonable to question whether such structural changes have consequential effects. Through professional training and repeated practice, experts develop stable schemas for judging quality that are not easily displaced by minor tweaks, such as rewording criteria or shifting emphasis, to formal rubrics. These cognitive frameworks become deeply ingrained professional practices, suggesting that structural changes may have limited influence on how experts assess project feasibility. When asked to evaluate feasibility alone rather than alongside other dimensions, experienced reviewers may simply apply their well-established mental models regardless of the evaluation structure. This expectation is reinforced by theories of expertise: expert judgment relies on pattern recognition, intuitive processing, and ingrained heuristics that operate largely independent of formal criteria (Kahneman, 2011). In the context of science in particular, professional identities are tied to stewarding multiple dimensions of quality simultaneously; reviewers see themselves as arbiters of novelty, feasibility, and broader contribution (Lamont, 2009). From this perspective, evaluation structures are primarily bureaucratic formalities: experts "know what good looks like," and the rubric mainly records judgments rather than shapes them.

Yet evaluation structure may represent a more consequential organizational lever than this expertise-based view suggests. We propose that multi-criteria evaluation creates a dual burden: it both

[2] https://grants.nih.gov/policy-and-compliance/policy-topics/peer-review/simplifying-review/framework

reduces the cognitive resources available for feasibility analysis and degrades how those limited resources get translated into judgments. These are not separate problems but interconnected symptoms of cognitive overload. Channeling attention through a single-criterion structure can reconfigure what evidence is noticed, how it is processed, and how it is weighted in feasibility judgments. Far from representing mere procedural formality, evaluation structure choices may fundamentally shape how organizations filter and direct innovative activity. The next section develops this claim by deriving testable hypotheses of how evaluation structures redirect attention and alter the quality of the feasibility signal.

How Single-Criterion Structures Reconfigure Feasibility Evaluation

Attention Allocation and Cognitive Load. Evaluation is fundamentally an attention-allocation problem. Because cognitive resources are finite, evaluators face systematic constraints on the depth and breadth of analysis they can conduct within limited review timeframes (Simon, 1955; Kahneman, 2011). With insufficient capacity to fully analyze all relevant information, evaluators must selectively direct their scarce attention across competing demands (Ocasio, 1997) or criteria (Kovács, Carnabuci, & Wezel, 2021). This allocation matters: what evaluators notice shapes what they value, determining which project attributes receive scrutiny, which problems surface, and ultimately which proposals gain support (Ocasio, 1997; Rhee & Leonardi, 2018). When organizations modify evaluation structures—shifting from multi-criteria to single-criterion assessment—they fundamentally restructure what aspects of a project receive scrutiny.

Multi-criteria evaluation imposes substantial cognitive load by requiring simultaneous assessment of novelty, impact, and feasibility. This load operates through two mechanisms. First, evaluators must distribute attention across multiple substantive criteria, analyzing feasibility while assessing scientific novelty and potential impact. This requires costly cognitive switching between distinct analytical frameworks (Monsell, 2003). Each dimension demands different expertise: novelty requires comparison to knowledge frontiers (Fleming, 2001; Kuhn, 1971), impact necessitates projecting future influence

(Kaplan & Vakili, 2015), and feasibility demands concrete assessment of implementation requirements (Boudreau et al., 2016). Second, evaluators allocate resources to maintaining coherence across dimensions—monitoring for inconsistencies between emerging judgments and managing the presentation of balanced assessment (Tetlock, 2002). This dual burden leaves feasibility with residual attention. Multi-criteria evaluators may scan for obvious implementation barriers but have reduced capacity for systematic interrogation across feasibility subcomponents such as budgets, timelines, technical dependencies, and resource constraints.

When evaluators shift to single-criterion assessment, they free substantial cognitive resources. Eliminating novelty and impact considerations removes both the cognitive switching between analytical frameworks and the effort required to maintain cross-dimensional coherence (Payne et al., 1993). Importantly, this redistribution occurs through structural design rather than individual choice—the evaluation structure channels attention by determining which dimensions are available for assessment. These liberated resources become available for redirection within the feasibility domain. Evaluators can now conduct more systematic feasibility analysis—examining budgets, timelines, technical requirements, regulatory constraints, and institutional barriers that previously received only superficial attention.

Yet the impact of this concentrated attention on feasibility is unclear. One possibility is that it yields more favorable feasibility assessments. When cognitive resources are freed from managing multiple criteria, evaluators can engage in more systematic solution search—actively exploring how apparent implementation challenges might be overcome (Nickerson & Zenger, 2004). Academic training specifically emphasizes solution generation alongside problem identification (Chi et al., 1988; Dane, 2010). Expert evaluators possess deep methodological knowledge and accumulated experience with how researchers adapt, substitute resources, and develop workarounds when facing constraints (Baker & Nelson, 2005). Under divided attention, this solution-oriented expertise may remain dormant, but concentrated feasibility focus could activate systematic exploration of implementation alternatives.

Single-criterion evaluation may also enable more sophisticated analogical reasoning—matching current implementation challenges to similar problems evaluators have seen successfully resolved in prior projects (Gentner & Markman, 1997; Holyoak & Thagard, 1995). Expert evaluators have observed how resourceful investigators overcome budget constraints, timeline pressures, and technical barriers across numerous projects. With sufficient cognitive bandwidth, they can systematically search their experience for relevant analogies, identifying precedents where similar challenges were successfully navigated.

Conversely, increased attention to feasibility may produce harsher assessments by surfacing previously overlooked implementation challenges. Systematic scrutiny across multiple feasibility dimensions—budgets, methods, timelines, and expertise requirements—increases the likelihood of identifying critical barriers. Each additional dimension examined represents another opportunity to identify disqualifying constraints. Moreover, evaluators possess broad expertise in recognizing common implementation failures but may lack project-specific knowledge needed to generate solutions. This asymmetry favors problem identification over solution generation, potentially leading to more critical assessments.

Why Critical Flaw Discovery Dominates in Scientific Evaluation. Although concentrated attention theoretically enables both problem discovery and solution search, we predict that single-criterion evaluation decreases feasibility scores because mechanisms favoring problem discovery overwhelm those supporting solution search. This dominance reflects both the structure of feasibility assessment and the professional context of scientific review. Feasibility assessment operates under "weakest link" logic where implementation barriers cannot be averaged away—a single critical flaw can render an entire project infeasible regardless of strengths in other areas (Åstebro & Elhedhli, 2006; Lane et al., 2022). This creates a fundamental asymmetry: each newly discovered problem has the potential to disqualify a project, while each solution generated merely addresses one concern among many. As evaluators systematically examine more implementation dimensions—budgets, timelines, design and technical requirements, regulatory constraints—the probability of encountering at least one disqualifying barrier increases. Unlike impact or novelty assessment, where multiple positive features can compensate for weaknesses, feasibility evaluation follows a conjunctive logic where all critical requirements must be satisfied simultaneously. A project cannot trade off an insufficient budget against a strong experimental design or compensate for missing regulatory approvals with an excellent research team. Each feasibility subcriterion operates as a veto point rather than a tradeable dimension.

The context of scientific evaluation amplifies this tendency through three reinforcing mechanisms. First, the ethos of organized skepticism trains scientists to systematically probe for weaknesses and potential failure modes (Merton, 1973). Professional incentives favor conservative assessment: the reputational cost of funding failures exceeds that of rejecting potentially successful projects, creating systematic bias toward identifying rather than dismissing implementation challenges (Lane et al., 2022). Second, feasibility assessment requires predicting future implementation success under uncertainty. Implementation barriers are often concrete and observable—budget shortfalls, missing expertise, underdeveloped or inappropriately applied technology, regulatory requirements—making them readily identifiable through systematic analysis. The inherent uncertainty of prediction means that deeper scrutiny systematically reveals more potential obstacles. Third, the knowledge differential between problem identification and solution generation favors critical assessment. Evaluators typically possess broad expertise in recognizing common implementation failures but lack the detailed project knowledge necessary to confidently assess how specific challenges might be overcome. This asymmetry means systematic analysis more readily surfaces concerns than generates confidence in implementability.

> **H1a (Feasibility Scrutiny Effect):** *Evaluators conducting single-criterion (feasibility-only) reviews will assign lower feasibility scores than evaluators conducting multi-criteria (novelty, impact, and feasibility) reviews.*

> **H1b (Problem Discovery Focus):** *Evaluators conducting single-criterion (feasibility-only) reviews will identify more feasibility-related problems than multi-criteria evaluators.*

Our prediction that systematic analysis leads to more critical feasibility assessment holds specifically for resource-constrained, time-limited evaluation of early-stage projects by scientific reviewers operating under professional norms of skepticism. Alternative contexts—evaluators with direct implementation experience, well-resourced settings, or extended evaluation periods—may shift this balance toward more optimistic feasibility judgments.

The shift from multi-criteria to single-criterion evaluation affects feasibility assessment through two complementary mechanisms. First, as described above, concentrated attention enables deeper scrutiny of implementation challenges. Second, structural elimination of novelty and impact dimensions removes cross-dimensional spillovers that can artificially inflate or deflate feasibility scores based on overall project impressions rather than implementation evidence.

Cross-Dimensional Spillovers in Feasibility Assessment

Multi-criteria evaluation creates systematic spillovers across assessment dimensions. When evaluators simultaneously judge novelty, impact, and feasibility, these assessments do not remain independent—judgments in one domain influence evaluations in others (Thorndike, 1920; Cooper, 1981). This interdependence reflects both cognitive and psychological processes: evaluators form holistic impressions that shape individual dimension scores, and they seek consistency across their judgments to maintain coherent overall assessments (Nisbett & Wilson, 1977). The result is a systematic correlation between feasibility scores and assessments of project promise, even when implementation viability should be evaluated independently.

On the one hand, positive spillovers may occur when high novelty and impact generate enthusiasm that extends to feasibility assessment. Evaluators impressed by a project's transformative potential may give it "the benefit of the doubt" on implementation challenges, rationalizing away concerns or imagining how creative investigators might overcome apparent barriers (Kahneman, 2011). Under this positive spillover mechanism, exciting projects receive inflated feasibility scores that reflect optimism about project promise rather than careful analysis of implementation requirements. Conversely, negative spillovers may dominate if evaluators perceive inherent tension between ambition and implementability. Novel, high-impact projects often employ untested methods, push technical boundaries, or require unconventional resources—features that may trigger skepticism about feasibility. Under negative spillovers, the very characteristics that make projects exciting also make them appear harder to implement, depressing feasibility scores below what implementation evidence alone would warrant.

The structural change from multi-criteria to single-criterion evaluation attenuates these spillover effects by removing novelty and impact from the evaluation structure. While feasibility-only evaluators still read complete proposals and may form impressions about project novelty or transformative potential, they are not required to formally evaluate or score these dimensions. This structural shift can redirect evaluators' analytical attention toward implementation evidence—budgets, timelines, technical requirements, institutional resources—and eliminates the pressure for cross-dimensional coherence that characterizes multi-criteria assessment. Without the need to integrate judgments across dimensions, evaluators' feasibility assessments become less influenced by their impressions of project promise.

Without the requirement to formally assess novelty and impact, evaluators face reduced pressure to maintain coherence across dimensions in their written evaluations. Rather than adjusting feasibility scores to align with overall project promise, evaluators can concentrate their analytical attention on implementation viability—the evidence directly relevant to their assigned task. This attentional focus fundamentally alters the cognitive process: instead of managing coherence across multiple formal judgments, evaluators direct their analytical resources toward feasibility-specific evidence. The single-criterion structure thus attenuates spillovers by reducing both the cognitive demand for cross-dimensional integration and the psychological pressure to maintain consistency across formal assessments.

This leads to the following empirical prediction: single-criterion evaluation should weaken the correlation between feasibility assessments and what would have been novelty/impact scores, regardless of whether that baseline correlation is positive or negative. By structurally limiting cross-dimensional comparison and redirecting attention toward implementation evidence, single-criterion structures reduce the spillover effects that create artificial dependencies between conceptually distinct dimensions.

> **H2 (Spillover Reduction):** *The correlation between feasibility scores and external assessments of project novelty/impact will be weaker under single-criterion (feasibility-only) evaluation than under multi-criteria evaluation.*

RESEARCH SETTING AND DESIGN

We collaborated with the translational science program at a large U.S. medical school to intervene in the evaluation process for a research pilot grant competition. Translation is the "process of turning laboratory, clinic, and community observations into interventions that improve the health of individuals and the public—from diagnostics and therapeutics to medical procedures and behavioral changes."[3] We chose this setting for three reasons: (i) the medical school's desire to enhance its evaluation process, (ii) the significance of feasibility evaluations in translational medicine, and (iii) the vital role of evaluation in grant allocations for academic research. The grant competition was open to any member with a university appointment pursuing an investigation related to human sensory systems in health or disease. Between October 19, 2021, and November 10, 2021, the competition received 47 complete proposals covering a variety of topics and awarded $50,000 to each of the top 10 proposals, totaling $500,000 for one year of funding.

Evaluator Recruitment and Selection

To aid in evaluator recruitment, at the end of the competition, the administrative team categorized the submitted proposals into six topic areas—sight, hearing, pain; smell/taste, touch; and proprioception/somatosensation—and using a university-wide database and Medical Subject Headings (MeSH) terms identified potential evaluators for each topic area. A subset of the top 100 most relevant

[3] https://ncats.nih.gov/translation.

evaluators within each topic area was randomly selected based on MeSH term search and invited to participate. In total, 97 evaluators accepted, resulting in 641 evaluator-proposal pairs across the six topics. Evaluators were randomly assigned to either the feasibility-only treatment (52 evaluators) or the multi-criteria control (45 evaluators) review condition. Each evaluator reviewed four to eight proposals within their area to allow multiple observations of evaluator behavior. Each proposal was reviewed by multiple evaluators in both conditions, ensuring that every proposal received independent assessments from evaluators assigned to both the multi-criteria and feasibility-only evaluation rubrics. On average, evaluators reviewed 6.47 proposals (s.d. = 1.93, min = 1, max = 9), and each proposal received a mean of 13.64 reviews (s.d. = 4.47, min = 7, max = 20).

Evaluator Instruction and Treatment

We implemented a triple-blind review process in which applicants were blinded to evaluators' identities, evaluators were blinded to applicants' identities and institutional affiliations, and evaluators were also blinded to the identities of other evaluators. All evaluations were conducted online using identical formatting and presentation. The full evaluation survey materials are provided in Online Appendix D.

Multi-Criteria Control Condition

Evaluators in the control condition conducted a multi-criteria review assessing three dimensions, each rated on a four-point scale:

1. *Impact* – is defined as "providing increased understanding of human sensory systems with potential translational benefit to human health assessed by patients, physicians, institutions, or society."
2. *Novelty* – is defined as "likely to lead to new technology or knowledge, an unanticipated application of existing technology, or a novel approach that enhances an established modality."
3. *Feasibility* – is defined as "addressing the proposed specific aims within the year of support using the proposed resources ($50K, direct costs, for 12 months)."

After assigning overall scores for each dimension, multi-criteria evaluators rated specific sub-criteria (i.e., components) related to novelty, impact, and feasibility, as indicated in Table 1. For the feasibility sub-criteria, the evaluators were asked to rate the extent to which they considered each component in their feasibility assessment (1 = Not at all, 4 = To a great extent). They also indicated their confidence in their feasibility ratings and provided open-ended comments on feasibility concerns and potential improvements. The instructions for the open-ended comments prompted evaluators to describe anticipated feasibility issues (“If you have concerns about project feasibility, please provide brief comments:”).

Table 1. Primary Evaluation Criteria and Sub-Criteria.

	Criteria		
	Novelty	**Impact**	**Feasibility**
Sub-Criteria	Produce new technology or knowledge	Patients	Experimental design
	Show unanticipated application of existing knowledge	Healthcare providers	Proposed analysis plan
	Generate a new approach for an established modality or concept	Healthcare systems	Proposed resources, including access to samples/patients
	Address a translational gap	Society	Alternative plan
	Improve an existing paradigm	Academic field	Timeframe (12 months)
			Scope – for $50K
			Regulatory requirements

Feasibility-Only Treatment Condition

The feasibility-only treatment condition differed from the multi-criteria control in two ways: evaluators received tailored instructions to assess feasibility alone, and only the feasibility rubric was displayed during their review. All other elements of the online review system—including proposal content, formatting, sequence, and platform functionality—were held constant across both conditions. Specifically, feasibility-only evaluators were instructed to focus solely on feasibility before beginning their reviews:

> *"[The grant organization] is interested in learning more about the scientific review process and is trying an alternative approach to reviewing proposals. For each of the proposals you have been assigned, we would like you to provide a focused review of their feasibility by answering the following questions. Please provide your review of the feasibility component in isolation, ignoring any considerations of the proposal's impact and innovation components."*

Evaluators in the feasibility-only treatment condition provided a mean of 6.38 reviews (s.d. = 2.03, min = 1, max = 9) per proposal. Each proposal received a mean of 7.19 (s.d. = 2.62, min = 3, max = 12) feasibility-only treatment reviews per proposal, for a total of 338 evaluator-proposal pairs. Evaluators in the multi-criteria control provided a mean of 6.59 reviews (s.d. = 1.82, min = 2, max = 8). Each proposal received a mean of 6.45 (s.d. = 2.22, min = 3, max = 10) multi-criteria control reviews per proposal for 303 evaluator-proposal pairs.

Main Variables

Dependent variables. The primary dependent variable, the *Feasibility score*, is the overall feasibility score evaluators assign to a proposal. Feasibility was assessed on a four-point scale (1 = Highly unlikely or not feasible, 2 = Possibly feasible, 3 = Likely feasible, and 4 = Fully feasible). We used the variable, *Number of feasibility sub-criteria considered,* to indicate the number of feasibility sub-criteria an evaluator considered *"to a great extent"* in terms of their contribution to a proposal's overall feasibility score.

Treatment variable. Our primary independent variable, *Feasibility-only treatment*, is a dummy variable indicating whether an evaluator was exogenously assigned to the feasibility-only treatment (1) or the multi-criteria control (0).

Independent variables. Although random assignment mitigates the need for controls, we included several variables to further investigate potential drivers of feasibility assessments. First, we measured evaluator topic area expertise based on the evaluator's self-reported familiarity with the sensory domain they were assigned to review on a scale of 1 ("Not at all familiar") to 4 ("Extremely familiar") and computed intellectual similarity by matching the proposal's MeSH terms (i.e., standardized biomedical indexing concepts used by the U.S. National Library of Medicine) to the MeSH terms in evaluators' first- and last-authored publications, yielding a cosine similarity score. We also tracked early career status (coded 1 for assistant professor or equivalent) and proposal characteristics (e.g., counts of figures, tables, references, and abstract word count). Additional procedural details concerning these variables are provided in the Online Appendix A.

Estimation Approach

We performed ordinary least squares (OLS) regressions to estimate the relationships between evaluators' *Feasibility scores* and the *Feasibility-only treatment* to investigate the effect of the feasibility-only evaluation rubric on the assigned feasibility scores. Our simplest model regresses *Feasibility scores* on the *Feasibility-only treatment* for each evaluator-proposal pair $\{i, j\}$. We then added evaluator covariates, X_i, (*Early career*, *Topic area expertise*), evaluator-proposal attributes (*Intellectual similarity*), ξ_{ij}, and proposal attributes (*Figures*, *Tables*, *References, Abstract*), γ_{ij}. The main model relating feasibility scores to the feasibility-only treatment is presented in Equation (1).

$$Feasibility\ scores_{ij} = \beta_0 + \beta_1 Feasibility\_Only\ Treatment_{ij}$$

$$+ \beta_3 X_i + \beta_4 \xi_{ij} + \beta_5 \gamma_{ij} + \varepsilon_{ij}. \ (1)$$

In all models, we cluster standard errors at the evaluator level. In certain model specifications, we added proposal fixed effects for all evaluators who reviewed more than one proposal. We performed robustness checks for equation (1) using ordinal logit models, in which we model the dependent variable as a categorical variable with four ordered possible values.

Human-LLM Content Coding Analysis of Qualitative Comments

We analyze evaluators' open-ended comments using a hybrid human-LLM coding approach to gain deeper insight into how the evaluation rubric shapes the conceptualization of feasibility. This approach begins with human coding to identify relevant codes and then scaling the analysis using OpenAI's GPT-4o—a state-of-the-art large language model capable of analyzing and coding text. This approach allowed us to capture how evaluators conceptualized feasibility in terms of interdependencies among sub-criteria and with novelty and impact.

Evaluators left comments on 307 of the 641 evaluations (47.9%). After removing brief, generic remarks such as "None" or "No concerns," we were left with 281 comments, corresponding to 43.8% of all evaluations. Responses ranged from terse one-liners (e.g., "The scope of the experiment is too ambitious given the budget") to multi-sentence paragraphs offering detailed critiques or advice.

Hybrid Human-LLM Coding Approach

Two authors independently analyzed a random sample of 57 comments distributed across treatment conditions (multi-criteria vs. feasibility-only). After excluding non-informative responses, we assessed 969 possible codes across feasibility, impact, and novelty sub-criteria (see Table 1). Through comparative analysis and resolution of coding discrepancies, we established a preliminary coding framework that captured anticipated and emergent themes within the data.

We then iteratively refined this initial framework through structured consultations with two program managers (scientists with more than ten years of grant evaluation experience) from the medical grant organization. These scientists reviewed our coding categories, suggested refinements, and helped identify domain-specific nuances crucial for accurate classification. Through multiple cycles of expert feedback and framework adjustment, we developed a consensus-based codebook incorporating predefined and emergent dimensions.

The finalized codebook then served as the foundation for developing LLM coding prompts, ensuring alignment between human-coded judgments and automated scaling of the full dataset. We used a few-shot prompting approach by providing the model with concrete examples illustrating each of the seven feasibility sub-criteria. Each comment underwent structured coding via a custom LLM prompt that instructed the model to (1) summarize the comment's main points and (2) identify mentions of the seven feasibility sub-criteria, novelty, and impact. The full prompt and details of the coding procedure are provided in Online Appendix C.

For example, consider the anonymized comment, which LLM coding classified as mentioning both experimental design and alternative plans:

> *The authors have identified numerous potential problems with this group of patients, and they have realistically made proposals for alternative plans. It may be challenging to keep the patients on task long enough to avoid fatigue, which could alter the [sensory] modality of the study since it follows the [task]. Even with breaking up the testing into additional sessions, subjects may forget portions of the prior [training], and repeat [training] may lead to fatigue before the auditory portion begins.*

To verify the alignment of LLM-based coding to human coders, two authors independently coded a random sample of 20 comments, stratified by treatment condition. The inter-rater reliability between human coders and the LLM was measured using Cohen's kappa, yielding an average $\kappa = 0.75$ across four human-LLM rating pairs. According to conventional thresholds (Landis & Koch, 1977), this value indicates substantial agreement between human and automated coding approaches. Further validation came from the two scientist-program managers from the grant funding program, who independently reviewed the LLM prompt and a separate sample of LLM-assigned codes and confirmed their accuracy. Detailed inter-rater reliability metrics are provided in Online Appendix C6.

RESULTS

Covariate balance checks indicate that the treatment and control assignments were successfully randomized, with no meaningful associations between the assignment and observed covariates (detailed summary statistics and balance checks are available in Online Appendix B1).

Hypothesis 1: Systematic Feasibility Analysis

We test Hypotheses 1a (feasibility scrutiny effect) and 1b (problem discovery mechanism) using two complementary approaches. To test Hypothesis 1a, we perform OLS models to estimate the average treatment effect of the feasibility-only evaluation rubric on feasibility scores (Table 2). To test Hypothesis 1b, we use both quantitative and qualitative analyses. First, we perform OLS models to estimate the average treatment effect of the feasibility-only evaluation rubric on the number of feasibility sub-criteria considered (Table 3). Then, we use the hybrid human-LLM analysis of the rich evaluator comment text.

Hypothesis 1a (Systematic analysis effect). In Table 2, Model 1 begins with the most basic comparison between the feasibility scores and feasibility-only treatment. Model 2 adds evaluator and evaluator-proposal level attributes, Model 3 adds proposal characteristics, and Model 4 adds proposal fixed effects. Finally, Model 5 investigates the interaction effects between the feasibility-only treatment and evaluator covariates.

Examining the relationships between feasibility scores and the feasibility-only treatment in Table 2, Model 1, we observe that the feasibility scores assigned by the feasibility-only treatment evaluators were 0.184 points lower, on average than those assigned by the multi-criteria control evaluators, corresponding to a 5.9 percent decrease in feasibility scores (Model 1: -0.184, $p < 0.05$). We observe that in Model 2, the coefficient for the *Feasibility-only treatment* remains stable and meaningful (Model 2: -0.186, $p < 0.05$). The coefficients for the *Feasibility-only treatment* also remain consistent in Models 3 and 4 (Model 3: -0.188, $p < 0.05$; Model 4: -0.197, $p < 0.05$), aligning with the randomization of evaluators into the multi-criteria versus feasibility-only conditions. Lastly, in Model 5, we include the interactions with the evaluator characteristics. The interaction terms indicate that the feasibility-only treatment effect is consistent across evaluator characteristics, including intellectual similarity to the proposal, topic expertise, and pre-tenure status. The estimated relationships in Table 2 indicate that the feasibility-only treatment condition leads to lower feasibility scores. Supplementary analyses in Online

Appendix B2 confirm that these results are robust to ordinal logit models. Thus, we find support for Hypothesis 1a.

Table 2. OLS Regression Models of Solution Feasibility Score on Feasibility-Only Treatment

	Dependent Variable: Feasibility Score				
	Model 1	Model 2	Model 3	Model 4	Model 5
Baseline = multi-criteria rubric					
Feasibility-only treatment	-0.184*	-0.186*	-0.188*	-0.197*	-0.495+
	(0.090)	(0.090)	(0.090)	(0.085)	(0.295)
Intellectual similarity		-0.003	-0.003	-0.005	-0.017
		(0.009)	(0.009)	(0.011)	(0.012)
Topic expertise		-0.009	-0.014	-0.055	-0.031
		(0.042)	(0.042)	(0.050)	(0.068)
Pre-tenure		0.063	0.063	0.089	0.024
		(0.101)	(0.102)	(0.100)	(0.169)
Figures			0.009		0.010
			(0.019)		(0.018)
Tables			0.007		0.005
			(0.052)		(0.052)
References			-0.001		-0.001
			(0.004)		(0.004)
Abstract			-0.002		-0.001
			(0.001)		(0.001)
Treatment × Intellectual similarity					0.030+
					(0.018)
Treatment × Topic expertise					0.034
					(0.088)
Treatment × Pre-tenure					0.092
					(0.205)
Intercept	3.116***	3.110***	3.477***		3.613***
	(0.069)	(0.168)	(0.310)		(0.324)
N	641	641	641	641	641
R2	0.013	0.014	0.017	0.163	0.022
FE: proposal				X	

Notes: standard errors clustered by evaluator. + $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Hypothesis 1b (Problem discovery focus). First, Table 3, Model 1 begins with the most basic comparison between the number of feasibility sub-criteria considered and the feasibility-only treatment. Model 2 adds evaluator and evaluator-proposal level attributes, Model 3 adds proposal characteristics, and Model 4 adds proposal fixed effects. Finally, Model 5 investigates the interaction effects between the feasibility-only treatment and evaluator covariates. In Model 1, we observe that the feasibility-only treatment evaluators considered, on average, 0.614 more sub-criteria than the multi-criteria control evaluators, corresponding to a 26 percentage point increase (Model 1: 0.614, $p < 0.10$). We observe that in Model 2, the coefficient for the *Feasibility-only treatment* remains stable (Model 2: 0.616, $p < 0.10$). The coefficients for the *Feasibility-only treatment* also remain consistent in Models 3 and 4 (Model 3: 0.620, $p < 0.10$; Model 4: 0.594, $p < 0.10$). Lastly, the interaction terms in Model 5 indicate that the treatment effect is strengthened among evaluators who are more intellectually similar to the proposals being evaluated. The results in Table 3, therefore, suggest that the lower scores observed in the feasibility-only treatment compared to the multi-criteria control condition in Table 2 may be attributed to the treated evaluators' consideration of additional feasibility sub-criteria.

Table 3. OLS Regression Models of Number of Feasibility Sub-Criteria on Feasibility-Only Treatment

Dependent Variable: Number of Feasibility Sub-Criteria Considered					
	Model 1	Model 2	Model 3	Model 4	Model 5
Baseline = multi-criteria rubric					
Feasibility-only treatment	0.614+	0.616+	0.620+	0.594+	-1.110
	(0.324)	(0.325)	(0.326)	(0.328)	(1.208)
Intellectual similarity		0.005	0.005	0.016	-0.059
		(0.030)	(0.031)	(0.042)	(0.037)
Topic expertise		-0.072	-0.035	-0.181	-0.123
		(0.177)	(0.180)	(0.191)	(0.249)
Pre-tenure		-0.243	-0.277	-0.235	-0.584
		(0.403)	(0.399)	(0.392)	(0.490)
Figures			-0.006		0.002
			(0.052)		(0.051)
Tables			0.003		-0.009
			(0.150)		(0.150)
References			0.027*		0.023+
			(0.012)		(0.012)
Abstract			-0.004		-0.003
			(0.003)		(0.003)
Treatment × Intellectual sim					0.134*
					(0.057)

Treatment × Topic expertise					0.196
					(0.345)
Treatment × Pre-tenure					0.695
					(0.778)
Intercept	2.353***	2.753***	2.947**		3.699***
	(0.229)	(0.643)	(0.996)		(1.029)
N	641	641	641	641	641
R2	0.018	0.020	0.028	0.117	0.044
FE: proposal				X	

Notes: standard errors clustered by evaluator. + p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

Turning to qualitative comments, we observe that evaluators in the feasibility-only condition were more likely to comment than those in the multi-criteria condition. Specifically, 183 (54.1%) of the 338 feasibility-only evaluators provided concerns, compared to 98 (32.3%) of the 303 multi-criteria evaluators, χ^2((1, N = 641) = 29.958, $p < 0.001$)). The length of open-ended comments did not meaningfully differ between conditions. Multi-criteria evaluators averaged 298 characters versus 275 characters for feasibility-only evaluators ($t(184) = 0.800$, $p = 0.425$). These descriptive statistics indicate that the feasibility-only rubric primarily increased the likelihood of identifying feasibility issues, not the length of responses.

Next, we turn to the results of the human-LLM hybrid approach of coding the feasibility comments into the distinct feasibility sub-criteria of experimental design, proposed analysis plan, timeframe, budget, resources, regulatory requirements, and alternative plan (see Table 1). Figure 1 reveals clear patterns in attention allocation between our two experimental conditions. First, we observe that the feasibility-only evaluators covered more feasibility sub-criteria in their comments. More specifically, feasibility-only evaluators mentioned five of the seven sub-criteria more frequently than in the multi-criteria condition, suggesting that the feasibility-only treatment reshaped not only the breadth of their attention on feasibility (i.e., more mentions of feasibility sub-criteria issues) but also how evaluators allocated their attention (i.e., more attention to "timeframe", "regulatory requirements", "alternative plan", and "experimental design", and less attention to "budget" and "resources").

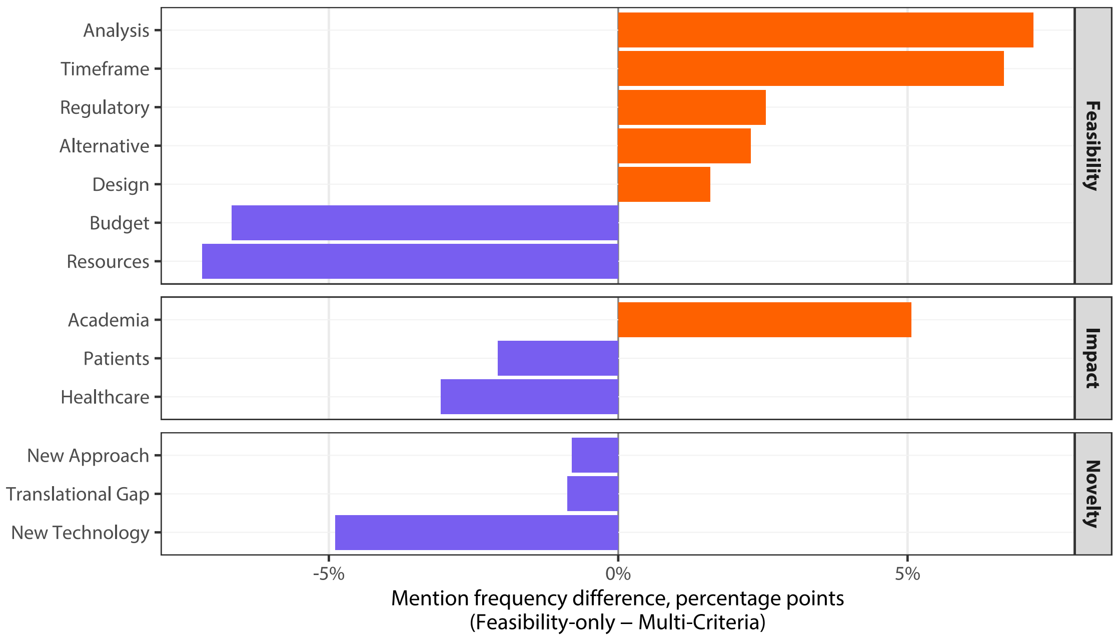

Figure 1. Difference in Sub-criteria Mentions between Feasibility-Only and Multi-Criteria Conditions normalized as percentage points of total mentions in each condition. Positive values (orange) indicate sub-criteria mentioned more frequently by feasibility-only evaluators; negative values (purple) indicate higher mention frequency by multi-criteria evaluators. Based on 183 feasibility-only and 98 multi-criteria evaluator comments.

Overall, the patterns in the quantitative and qualitative analyses support our theoretical argument that systematic attention may lead to a more in-depth consideration of feasibility sub-criteria, revealing implementation challenges that remain hidden under a divided cognitive load. Thus, we find support for Hypothesis 1b, that evaluators conducting feasibility-only reviews will identify more feasibility-related problems than multi-criteria evaluators.

Hypothesis 2: Cross-Dimensional Spillovers

To test whether single-criterion evaluation mitigates spillover effects between feasibility and other project dimensions, we examine how feasibility scores correlate with novelty and impact assessments across evaluation formats. This investigation requires a proposal-level analysis because single-criterion evaluators assessed only feasibility, not novelty or impact. We therefore calculate mean feasibility scores for each proposal under both conditions, alongside mean novelty and impact scores from multi-criteria evaluations.

Figure 2 provides a visual overview of how proposal-level feasibility scores relate to novelty and impact assessments. As the figure shows, feasibility is positively related to novelty and impact in multi-criteria evaluations, but these associations disappear under single-criterion evaluations.

Table **4** reports the regression models that formally test their significance. Models 1-3 use feasibility scores from multi-criteria evaluators as dependent variables; Models 4-6 use feasibility scores from single-criterion evaluators.

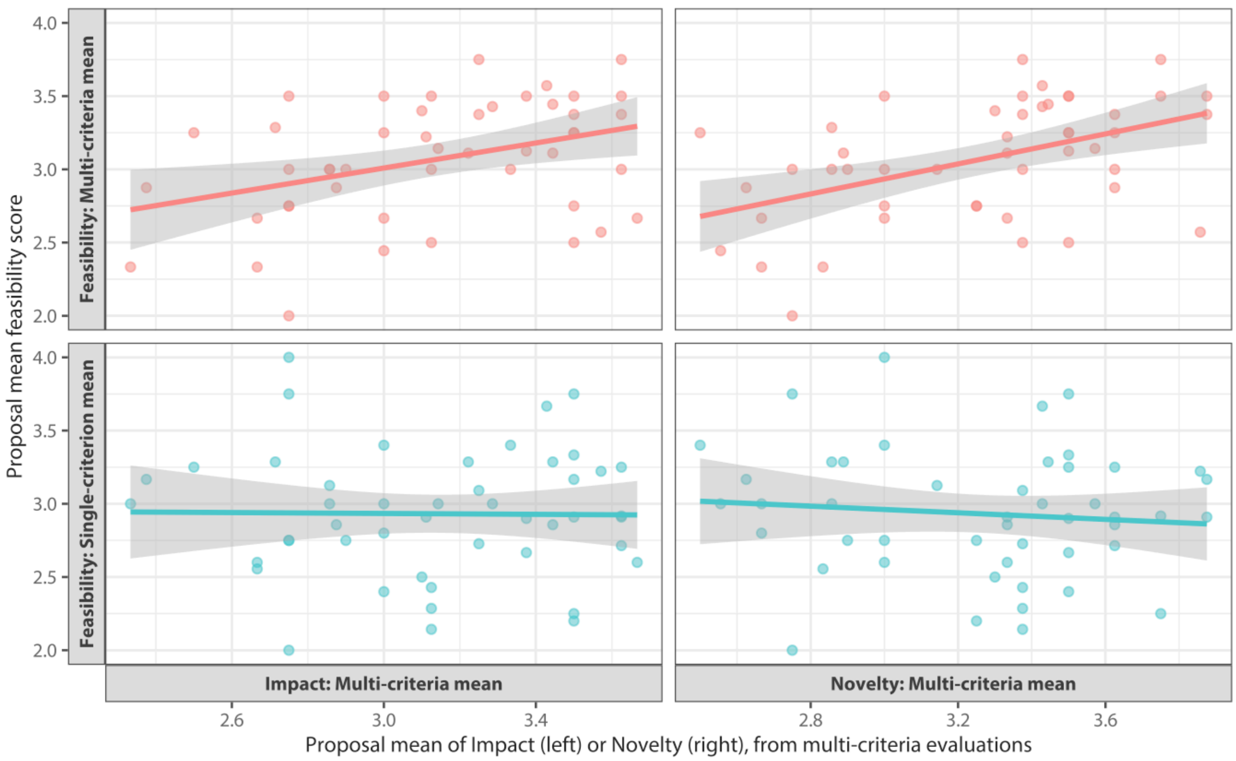

Figure 2. Cross-Dimensional Spillovers Emerge in Multi-Criteria but Not Single-Criterion Evaluations. Each point represents one proposal. The x-axes show mean Impact (left) or Novelty (right) scores from multi-criteria evaluators. The y-axes show mean Feasibility scores from multi-criteria evaluators (top row) and single-criterion evaluators (bottom row). Lines indicate OLS fits with 95% confidence intervals.

Table 4. OLS Regression Models of Feasibility on Novelty and Impact by Evaluation Format.

	Dependent Variable: Multi-Criteria Feasibility Score			*Dependent Variable: Single-Criterion Feasibility Score*		
	(1)	(2)	(3)	(4)	(5)	(6)
Novelty	0.512***		0.438*	-0.113		-0.201
	(0.142)		(0.200)	(0.172)		(0.243)
Impact		0.428**	0.110		-0.015	0.131

		(0.154)	(0.207)		(0.179)	(0.252)
Intercept	1.398**	1.726***	1.295*	3.300***	2.979***	3.177***
	(0.465)	(0.486)	(0.507)	(0.565)	(0.566)	(0.616)
N	47	47	47	47	47	47
R2	0.225	0.146	0.230	0.009	0.000	0.015

Notes: Novelty and impact scores represent mean assessments from multi-criteria evaluators for each proposal. + p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

In the multi-criteria condition (Table **4**, Models 1-3), feasibility scores correlate positively and significantly with both novelty (Model 1, $\beta = 0.512$, $p < 0.001$) and impact (Model 2, $\beta = 0.428$, $p < 0.01$) when examined separately. When both dimensions are included together (Model 3), novelty remains significant ($\beta = 0.438$, $p < 0.05$) while impact becomes non-significant ($\beta = 0.110$, $p = 0.598$), suggesting novelty drives the primary spillover effect when controlling for impact. These correlations explain substantial variance in multi-criteria feasibility assessments (Model 3, $R^2 = 0.230$).

The pattern reverses in the single-criterion condition (Table **4**, Models 4-6). These models reveal no significant relationships between feasibility scores and either novelty (Model 4, $\beta = -0.113$, $p = 0.52$) or impact (Model 5, $\beta = -0.015$, $p = 0.93$), whether examined separately or jointly (Model 6). The coefficients not only lose statistical significance but shift directionally—the novelty coefficient even reverses from positive to negative, though remaining non-significant. Furthermore, the explained variance of novelty and impact on feasibility drops ten-fold, to near-zero (Model 6, $R^2 = 0.015$).

This attenuation provides support for H2. Multi-criteria evaluation creates systematic spillovers where projects perceived as novel and impactful receive elevated feasibility scores. Keeping impact constant, a 1-point increase in novelty associates with a 0.44-point increase in feasibility. We do not find evidence of these coherence pressures in single-criterion evaluations, suggesting that they successfully redirect evaluators' attention exclusively to implementation viability. The structural removal of novelty and impact from the evaluation task essentially negates the influence of impact and novelty on feasibility judgments.

Robustness Check: Attention to Dimensions in Written Comments

The proposal-level correlation analysis faces the limitation that it compares assessments across evaluator groups. As a more direct test, we turn to within-evaluator evidence by analyzing the written comments evaluators provided when asked to discuss feasibility concerns. If multi-criteria feasibility judgments are influenced by novelty and impact, we should expect them to reference those dimensions more often in their feasibility comments than single-criterion evaluators do. To assess this, we apply our LLM-based coding of evaluator text to compare the frequency with which novelty and impact are mentioned across the two conditions.

Returning to Figure 1, we observe that multi-criteria evaluators were more likely to mention all novelty subcriteria and all but one of the impact subcriteria (academic impact) when explaining feasibility. This consistent pattern is aligned with cross-dimensional spillovers: when evaluating multiple criteria, novelty and impact considerations seep into feasibility reasoning. While academic impact was somewhat more frequently mentioned by feasibility-only evaluators, our regression results already showed that impact played a limited role in explaining spillovers once novelty was accounted for (Table **4**, Model 3), suggesting that this exception was unlikely to substantively shape feasibility assessments. Taken together with the proposal-level correlations, these results provide convergent evidence across and within evaluators that multi-criteria evaluation intertwines dimensions, whereas single-criterion evaluation keeps feasibility reasoning more insulated.

DISCUSSION

This study examines how the structure of evaluation rubrics influences feasibility judgments in innovation assessment. Across many organizational settings, multi-criteria evaluation is intended to promote comprehensive and balanced decision-making. Yet our experiment shows that this widely used approach can unintentionally shape evaluators' assessments of project feasibility. Specifically, we find that multi-criteria rubrics—by requiring simultaneous assessment of novelty, impact, and feasibility—alter the cognitive processes underlying feasibility judgment through two reinforcing mechanisms. When evaluators consider multiple dimensions at once, divided attention constrains systematic problem discovery, while cross-dimensional spillovers lead feasibility assessments to mirror overall project promise rather than implementation evidence. These patterns suggest that organizations may sometimes advance projects whose implementation challenges remain insufficiently examined—a dynamic that may contribute to the enduring puzzle of why novel, high-impact initiatives frequently encounter feasibility problems despite passing rigorous evaluation.

These mechanisms operate through distinct yet complementary pathways. The attention mechanism reflects fundamental cognitive constraints: evaluators assessing multiple dimensions must allocate limited cognitive resources, leaving feasibility with less attention for detailed scrutiny. Our qualitative analysis showed that feasibility-only evaluators examined a broader range of implementation issues—experimental design, analysis plans, timeline realism, regulatory requirements, and contingency planning—whereas multi-criteria evaluators focused more narrowly on budgets and resources, dimensions that are simpler to verify but less diagnostic of overall feasibility. The spillover mechanism captures a subtler dynamic: evaluators' impressions of a project's novelty and impact systematically influence their feasibility judgments, consistent with coherence-seeking tendencies that promote internally consistent evaluations across criteria. Together, these processes indicate that the design of evaluation rubrics plays a consequential role in shaping how implementation risks are recognized.

Theoretical and Methodological Contributions

This research contributes to innovation management theory in several ways. First, our study enriches understanding of perceived risk and reward in innovation (Ferguson & Carnabuci, 2017; Fleming, 2001; Khanna et al., 2016; Scherer & Harhoff, 2000) and underscores the critical importance of prospective evaluations in shaping perspectives on the likelihood of success or failure of novel projects. In particular, our results in

Table **4** demonstrate that both novelty and impact are correlated with feasibility scores in the multi-criteria condition; however, the relationships are stronger between novelty and feasibility than between impact and feasibility. Unlike previous studies that have relied mostly on retrospective analyses of innovative outcomes and concluded that novel projects bear greater risk and are more likely to fail, our approach has led us to propose that outcomes might be influenced by the evaluation process itself. In other words, single-criterion evaluation rubrics that focus on feasibility, for instance, may afford better management and possibly even the elimination of some of the perceived risks and failures associated with novel, high-impact projects. When combined with multi-criteria approaches, focused feasibility assessment can improve project success by enabling earlier detection and mitigation of implementation risks.

Second, our findings reveal a fundamental limitation of alternative scoring schemes within standard multi-criteria rubrics. Even sophisticated approaches—such as selecting projects with the fewest weaknesses, highest consensus scores, or specific feasibility thresholds—cannot overcome the systematic spillovers we document between feasibility assessments and evaluations of novelty and impact. Our H2 results demonstrate that multi-criteria evaluation creates structural dependencies between conceptually distinct dimensions: evaluators' feasibility judgments become entangled with their impressions of project novelty and impact, through both positive and negative spillovers.

By contrast, single-criterion feasibility evaluation eliminates these spillovers structurally rather than attempting to correct for them statistically. When evaluators are not asked to form impressions about novelty or impact (because these dimensions are absent from their task), their feasibility scores become more rooted in implementation evidence—budgets, timelines, technical requirements, and resource constraints—rather than being colored by impressions of the project's potential. Our qualitative analyses confirm this redirection of attention: feasibility-only evaluators systematically examined more feasibility sub-criteria and discovered implementation challenges that remained hidden under divided cognitive load. This suggests that assessments focused solely on feasibility enable evaluators to discern subtler implementation barriers that emerge only through thorough information processing and concentrated attention (Ocasio, 1997).

Third, our study advances understanding of how organizational design shapes cognition by revealing that seemingly minor procedural choices function as powerful cognitive interventions. While attention-based theories emphasize that organizational structures channel what decision-makers notice (Ocasio, 1997; Joseph & Wilson, 2018; Kovács et al., 2021), we demonstrate that evaluation rubrics do more than redistribute attention—they fundamentally reconfigure the cognitive processes evaluators employ. Single-criterion evaluation does more than give evaluators “more time” for feasibility; it eliminates the cognitive overhead of maintaining coherence across dimensions, removes cross-dimensional spillovers, and enables qualitatively different analytical strategies (systematic disaggregation of sub-criteria rather than holistic impressions). This shifts the theoretical focus from attention as a scarce resource to be allocated, toward understanding how organizational procedures actively construct the cognitive structures within which evaluators make their assessments.

Fourth, this study makes a methodological contribution by demonstrating how field experiments combined with human-LLM analysis can reveal cognitive mechanisms underlying organizational processes. While cognitive processes are typically difficult to observe directly, our approach shows how careful research design and novel analytical methods can unpack these mechanisms at scale. This methodological innovation opens new possibilities for studying how organizational structures affect individual cognition and decision-making in practice.

Limitations and Future Research

Our experimental design establishes that evaluation structure systematically affects feasibility assessment through two mechanisms: attention allocation and cross-dimensional spillover effects. These mechanisms are theoretically general, but their magnitudes and consequences likely vary across contexts and merit careful consideration. The direction of spillover depends on how evaluators perceive the relationship between dimensions. In our translational medicine setting, we observe positive spillovers: novel, high-impact projects receive higher feasibility scores, potentially leading organizations to approve transformative but infeasible initiatives. However, in domains where novel approaches are perceived as inherently riskier or less proven, the correlation might reverse: evaluators might systematically deflate feasibility scores for transformative projects, leading to false negatives rather than false positives. Similarly, spillovers may be attenuated when evaluators face stronger accountability (e.g., venture capitalists investing their own capital) (Bian et al., 2021) or when feasibility criteria are more objectively verifiable. The optimal evaluation structure thus depends on whether false positives (funding infeasible projects) or false negatives (rejecting feasible innovations) pose greater organizational risk (Krieger et al., 2022; Lane et al., 2022).

Future research examining alternative institutional contexts—including scientific fields with varying degrees of theoretical consolidation and openness to unconventional approaches (Azoulay et al., 2019), corporate R&D and new product development settings (Criscuolo et al., 2021; Katila & Ahuja, 2002), and contexts with different organizational costs of false positives versus false negatives—would help map the boundary conditions and identify interventions for improving evaluation processes.

Our field experimental design tests the core question of whether a multi-criteria versus a single-criterion evaluation structure systematically shapes feasibility assessment. The practical constraints of conducting field experiments with expert evaluators in consequential funding decisions precluded testing additional conditions, such as sequential evaluation structures for different criteria, systematic manipulation of proposal load, or isolated assessment of impact or novelty alone. Our design establishes that evaluation structure shapes feasibility assessment in realistic organizational settings; what we cannot determine is the full range of structural variations that might attenuate or amplify these effects. Controlled laboratory experiments with simulated proposals could provide additional manipulations to isolate specific mechanisms, though at the potential cost of ecological validity.

An alternative explanation for our findings merits attention: whether our findings reflect coherence pressure or simply cognitive workload differences. Multi-criteria evaluators faced inherently higher cognitive demands—they needed to assess three distinct dimensions (impact, novelty, and feasibility) rather than just one, requiring mental task-switching and distributing limited cognitive resources across multiple evaluation criteria. Consequently, their less thorough feasibility assessments might stem partly from having less time and attention to devote to feasibility after addressing other dimensions rather than from fundamentally different cognitive processing strategies. We deliberately did not constrain evaluation time in our design, allowing evaluators to allocate attention as they would in naturalistic review settings. However, our qualitative analyses suggest systematic differences in how evaluators conceptualize feasibility across conditions, indicating that evaluation structure shapes whether evaluators integrate feasibility with other dimensions rather than merely constraining available attention. Future research could more precisely isolate these mechanisms through designs that control for total evaluation time and cognitive load. For instance, researchers could employ matched-effort conditions where single-criterion evaluators complete additional unrelated tasks to equalize total cognitive effort, allowing more direct comparison of cognitive processing strategies independent of workload differences.

CONCLUSION

Evaluation rubrics are not neutral instruments—they are cognitive interventions that determine what evaluators notice and how they judge evidence. This research reveals a fundamental tension in innovation evaluation: comprehensiveness may undermine precision. Multi-criteria evaluation, designed to capture innovation's multidimensional nature, can inadvertently compromise feasibility assessment through fragmented attention and cross-dimensional spillover. The path forward is not more comprehensive evaluation but more strategic evaluation—assessing critical dimensions independently rather than simultaneously. Organizations must recognize that how they structure evaluation is as consequential as what they evaluate, shaping not just which innovations are funded but which implementation barriers are detected.

REFERENCES

Abernathy, W. J., & Clark, K. B. (1985). Innovation: Mapping the winds of creative destruction. *Research Policy*, *14*(1), 3–22.

Amabile, T. M. (1983). Brilliant but cruel: Perceptions of negative evaluators. *Journal of Experimental Social Psychology*, *19*(2), 146–156.

Åstebro, T., & Elhedhli, S. (2006). The effectiveness of simple decision heuristics: Forecasting commercial success for early-stage ventures. *Management Science*, *52*(3), 395–409.

Azoulay, P., Fons-Rosen, C., & Graff Zivin, J. S. (2019). Does science advance one funeral at a time? *American Economic Review*, *109*(8), 2889–2920.

Azoulay, P., & Li, D. (2020). *Scientific grant funding* (Vol. 26889). Cambridge, MA: National Bureau of Economic Research.

Baer, M. (2012). Putting creativity to work: The implementation of creative ideas in organizations. *Academy of Management Journal*, *55*(5), 1102–1119.

Baer, M., Dirks, K. T., & Nickerson, J. A. (2013). Microfoundations of strategic problem formulation. *Strategic Management Journal*, *34*(2), 197–214.

Bamberger, P. A. (2019). On the replicability of abductive research in management and organizations: Internal replication and its alternatives. *Academy of Management Discoveries*, *5*(2), 103-108.

Baker, T. and Nelson, R.E. (2005). Creating something from nothing: Resource construction through entrepreneurial bricolage. Administrative Science Quarterly, 50(3), 329–366.

Behfar, K., & Okhuysen, G. A. (2018). Perspective—Discovery within validation logic: Deliberately surfacing, complementing, and substituting abductive reasoning in hypothetico-deductive inquiry. *Organization Science*, *29*(2), 323-340.

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

Bian, J., Greenberg, J., Li, J., & Wang, Y. (2021). Good to go first? Position effects in expert evaluation of early-stage ventures. *Management Science*, *68*(1), 300–315.

Bikard, M., & Marx, M. (2020). Bridging academia and industry: How geographic hubs connect university science and corporate technology. *Management Science*, *66*(8), 3425–3443.

Boudreau, K. J., Guinan, E. C., Lakhani, K. R., & Riedl, C. (2016). Looking across and looking beyond the knowledge frontier: Intellectual distance, novelty, and resource allocation in science. *Management Science*, *62*(10), 2765–2783.

Briel, M., Olu, K. K., Von Elm, E., Kasenda, B., Alturki, R., Agarwal, A., Bhatnagar, N., & Schandelmaier, S. (2016). A systematic review of discontinued trials suggested that most reasons for recruitment failure were preventable. *Journal of Clinical Epidemiology*, *80*, 8–15.

Carnabuci, G., & Operti, E. (2013). Where do firms' recombinant capabilities come from? Intraorganizational networks, knowledge, and firms' ability to innovate through technological recombination. *Strategic Management Journal*, 34(13), 1591–1613.

Chang, L.W., Kirgios, E.L., Mullainathan, S. and Milkman, K.L., 2024. Does counting change what counts? Quantification fixation biases decision-making. Proceedings of the National Academy of Sciences, 121(46).

Criscuolo, P., Dahlander, L., Grohsjean, T., & Salter, A. (2017). Evaluating novelty: The role of panels in the selection of R&D projects. *Academy of Management Journal*, *60*(2), 433–460.

Criscuolo, P., Dahlander, L., Grohsjean, T., & Salter, A. (2021). The sequence effect in panel decisions: Evidence from the evaluation of research and development projects. *Organization Science*, *32*(4), 987–1008.
Dougherty, D. (1992). A practice-centered model of organizational renewal through product innovation. *Strategic Management Journal*, *13*(S1), 77–92.
Eklund, J., Raj, M., & Eggers, J. P. (2025). Attention focus and new opportunities: The moderating role of managerial attention to alternative issues. *Organization Science*, *36*(1), 21-39.
Falchetti, D., Cattani, G., & Ferriani, S. (2022). Start with "why," but only if you have to: The strategic framing of novel ideas across different audiences. *Strategic Management Journal*, *43*(1), 130-159.
Ferguson, J. P., & Carnabuci, G. (2017). Risky recombinations: Institutional gatekeeping in the innovation process. *Organization Science*, *28*(1), 133-151.
Fleming, L. (2001). Recombinant uncertainty in technological search. *Management Science*, *47*(1), 117–132.
Fleming, L., & Sorenson, O. (2004). Science as a map in technological search. *Strategic Management Journal*, *25*(8–9), 909–928.
Fogel, D. B. (2018). Factors associated with clinical trials that fail and opportunities for improving the likelihood of success: A review. *Contemporary Clinical Trials Communications*, *11*, 156–164.
Franzoni, C., & Stephan, P. (2023). Uncertainty and risk-taking in science: Meaning, measurement and management in peer review of research proposals. *Research Policy*, *Research Policy 52*(3), 104706.
Guler, I. (2007). Throwing good money after bad? Political and institutional influences on sequential decision making in the venture capital industry. *Administrative Science Quarterly*, *52*(2), 248–285.
Holyoak, K. J., & Simon, D. (1999). Bidirectional reasoning in decision making by constraint satisfaction. *Journal of Experimental Psychology: General*, *128*(1), 3.
Hsee, C.K., Loewenstein, G.F., Blount, S. and Bazerman, M.H. (1999). Preference reversals between joint and separate evaluations of options: A review and theoretical analysis. Psychological Bulletin, 125(5), 576–590.
Joseph, J., & Wilson, A. J. (2018). The growth of the firm: An attention-based view. *Strategic Management Journal*, *39*(6), 1779-1800.
Joseph, J., & Ocasio, W. (2012). Architecture, attention, and adaptation in the multibusiness firm: General electric from 1951 to 2001. *Strategic Management Journal*, *33*(6), 633-660.
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.
Kaplan, S., & Vakili, K. (2015). The double-edged sword of recombination in breakthrough innovation. *Strategic Management Journal, 36*(10), 1435–1457.
Kerr, W. R., Nanda, R., & Rhodes-Kropf, M. (2014). Entrepreneurship as experimentation. *Journal of Economic Perspectives*, *28*(3), 25–48.
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.
Knudsen, T., & Levinthal, D. A. (2007). Two faces of search: Alternative generation and alternative evaluation. *Organization Science*, *18*(1), 39–54.
Kovács, B., Carnabuci, G., & Wezel, F. C. (2021). Categories, attention, and the impact of inventions. *Strategic Management Journal*, 42(5), 992–1023.
Krieger, J., Li, D., & Papanikolaou, D. (2022). Missing novelty in drug development. *The Review of Financial Studies*, *35*(2), 636-679.

Kuhn, T. S. (1977). *The Essential Tension: Selected Studies in Scientific Tradition and Change*. Chicago: University of Chicago Press.

Kuhn, T. S. (2003). Objectivity, value judgment, and theory choice. In A. Bird, & J. Ladyman (Eds.), *Arguing about Science* (pp. 74–86). New York: Routledge.

Lamont, M. (2009). *How Professors Think: Inside the Curious World of Academic Judgment*. Harvard University Press.

Landis, J. R., & Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. *Biometrics*, *33*(1), 159–174.

Lane, J., Ganguli, I., Gaule, P., Guinan, E., & Lakhani, K. R. (2021). Engineering serendipity: When does knowledge sharing lead to knowledge production? *Strategic Management Journal*, *42*(6).

Lane, J., Teplitskiy, M., Gray, G., Ranu, H., Menietti, M., Guinan, E., & Lakhani, K. R. (2022). Conservatism gets funded? A field experiment on the role of negative information in novel project evaluation. *Management Science*, *68*(6), 4478–4495.

Leiponen, A., & Helfat, C. E. (2010). Innovation objectives, knowledge sources, and the benefits of breadth. *Strategic Management Journal*, *31*(2), 224–236.

Leonardi, P. M., Neeley, T. B., & Gerber, E. M. (2012). How managers use multiple media: Discrepant events, power, and timing in redundant communication. Organization Science, 23(1), 98-117.

Levinthal, D. A., & March, J. G. (1993). The myopia of learning. *Strategic Management Journal*, *14*(2), 95-112.

Li, D., 2017. Expertise versus Bias in Evaluation: Evidence from the NIH. *American Economic Journal: Applied Economics*, *9*(2), 60–92.

March, J. G., & Simon, H. A. (1958). *Organizations*. Wiley.

Mount, M. P., Baer, M., & Lupoli, M. J. (2021). Quantum leaps or baby steps? Expertise distance, construal level, and the propensity to invest in novel technological ideas. *Strategic Management Journal*, *42*(8), 1490–1515.

NIH Research Grants – Digital Media Kit. (2024, December 19). National Institutes of Health; National Institutes of Health. https://www.nih.gov/news-events/nih-research-grants-digital-press-kit.

Ocasio, W. (1997). Towards an attention-based view of the firm. *Strategic Management Journal*, *18*(S1), 187–206.

Ocasio, W. (2011). Attention to attention. *Organization Science*, *22*(5), 1286–1296.

Prahalad, C. K., & Hart, S. L. (2002). The fortune at the bottom of the pyramid. Strategy+Business, 26.

Rhee, L., & Leonardi, P. M. (2018). Which pathway to good ideas? An attention-based view of innovation in social networks. *Strategic Management Journal*, *39*(4), 1188–1215.

Rosenberg, N., & Nelson, R. R. (1994). American universities and technical advance in industry. *Research Policy*, *23*(3), 323–348.

Scherer, F. M., & Harhoff, D. (2000). Technology policy for a world of skew-distributed outcomes. *Research Policy*, *29*(4–5), 559–566.

Schumpeter, J. A. (1942). *Capitalism, Socialism and Democracy*. Routledge.

Simon, H. A. (1955). A behavioral model of rational choice. *Quarterly Journal of Economics, 69*(1), 99–118.Sleesman, D. J., Lennard, A. C., McNamara, G., & Conlon, D. E. (2018). Putting escalation of commitment in context: A multilevel review and analysis. *Academy of Management Annals*, *12*(1), 178–207.

Stephan, P. (2012). *How Economics Shapes Science*. Cambridge, MA: Harvard University Press.

Tan, D. (2022). The road not taken: Technological uncertainty and the evaluation of innovations. *Organization Science*, *34*(1), 156–175.

Teodoridis, F., Bikard, M., & Vakili, K. (2019). Creativity at the knowledge frontier: The impact of specialization in fast-and slow-paced domains. *Administrative Science Quarterly*, *64*(4), 894-927.

Tripsas, M., & Gavetti, G. (2000). Capabilities, cognition, and inertia: Evidence from digital imaging. *Strategic Management Journal*, *21*(10-11), 1147–1161.

Wang, J., Veugelers, R., & Stephan, P. (2017). Bias against novelty in science: A cautionary tale for users of bibliometric indicators. *Research Policy*, *46*(8), 1416–1436.

Weitzman, M. L. (1998). Recombinant growth. *Quarterly Journal of Economics*, *113*(2), 331–360.Zhou, K. Z., & Wu, F. (2010). Technological capability, strategic flexibility, and product innovation. *Strategic Management Journal, 31*(5), 547–56.