Working Paper 22-071

Are Experts Blinded by Feasibility? Experimental Evidence from a NASA Robotics Challenge

Jacqueline N. Lane
Zoe Szajnfarber
Jason Crusan
Michael Menietti
Karim R. Lakhani

Are Experts Blinded by Feasibility? Experimental Evidence from a NASA Robotics Challenge

Jacqueline N. Lane
Harvard Business School and Laboratory for Innovation Science at Harvard

Zoe Szajnfarber
George Washington University

Jason Crusan
George Washington University

Michael Menietti
Harvard Business School and Laboratory for Innovation Science at Harvard

Karim R. Lakhani
Harvard Business School and Laboratory for Innovation Science at Harvard

Working Paper 22-071

Funding for this research was provided in part by Harvard Business School, NASA Cooperative Agreement #NNX16AC77A, and NSF award CMMI-1535539.

Are Experts Blinded by Feasibility?

Experimental Evidence from a NASA Robotics Challenge

Jacqueline N. Lane[1*], Zoe Szajnfarber[2], Jason Crusan[2], Michael Menietti[1] and Karim R. Lakhani[1]

[1]Harvard Business School and Laboratory for Innovation Science at Harvard,

[2]George Washington University

[*]Correspondence to: jnlane@hbs.edu

Abstract

Resource allocation decisions play a dominant role in shaping a firm's technological trajectory and competitive advantage. Recent work indicates that innovative firms and scientific institutions tend to exhibit an anti-novelty bias when evaluating new projects and ideas. In this paper, we focus on shedding light into this observed pattern by examining how evaluator expertise in the problem's focal domain shapes the relationship between novelty and feasibility in evaluations of quality for technical solutions. To estimate relationships, we partnered with NASA and Freelancer.com, an online labor marketplace, to design an evaluation challenge, where we recruited 374 evaluators from inside and outside the technical domain to rate 101 solutions drawn from nine robotics challenges. This resulted in 3,869 evaluator-solution pairs, in which evaluators were randomly assigned to solutions to facilitate experimental comparisons. Our experimental findings, complemented with text analysis of the evaluators' comments, indicate that domain experts exhibit a *feasibility preference*, focusing first on the feasibility of a solution as the primary indicator of its quality, while discounting riskier but more novel solutions. This results in a tradeoff in which highly feasible but less novel solutions are judged as being higher in quality, shedding light into why experts prefer more incremental ideas over more radical but untested ideas.

Keywords: evaluations, resource allocation, novelty, feasibility, technological innovation, field experiment

1. Introduction

Ever since Schumpeter (1942)'s essential work on the theory of creative destruction, strategy scholars have long viewed organizational innovation and renewal as key to firm survival, economic growth, and competitive advantage in dynamic environments (Benner and Tushman 2003, Danneels 2002, Nelson and Winter 1982, Schumpeter 1942, Teece et al. 1997, Tushman and Anderson 1986). A firm's capacity to expand its organizational competences over time through new products and processes depends on its ability to both exploit current technologies and resources to increase productivity as well as create variation through exploratory innovation (March 1991, Rosenberg 1972, Teece et al. 1997). As firms learn through repetition by exploiting their current capabilities to gain greater efficiency, their innovation becomes increasingly incremental, characterized by small changes in their technological trajectory (Benner and Tushman 2003, Cyert and March 1963, Henderson 1993). In contrast, radical innovation draws on novel scientific and engineering principles that instigate fundamental changes in firms' technological trajectories and competences to make major technical advances (Abernathy and Clark 1985, Benner and Tushman 2003, Dosi 1982, Levinthal and March 1993).

Radical innovation is challenging for established firms, because it requires different organizational capabilities that can challenge and even destroy existing capabilities, assets and experience (Abernathy and Utterback 1978, Eisenhardt and Martin 2000, Henderson and Clark 1990, Leonard-Barton 1992, Tripsas 1997). Many established firms fail during periods of rapid technological change, often due to the inconsistencies between activities focused on productivity and cost improvements and those emphasizing radical innovation and flexibility (Abernathy 1976, Benner and Tushman 2003, Burns and Stalker 1961, Gilbert and Newbery 1982). This strongly suggests that a firm's patterns of resource allocation influences the types of innovations that are evidenced from their research and development (R&D) pipelines (Christensen and Bower 1996, Teece et al. 1997).

A fundamental challenge firms face is how to effectively allocate resources across alternative projects (Bower 1972, Maritan and Lee 2017, Noda and Bower 1996). Critical to resource allocation decisions is the evaluation and selection of high quality projects from a firm's R&D pipeline (Åstebro and

Elhedhli 2006, Boudreau et al. 2016). These processes often rest in the hands of innovation managers and executives who draw on the judgments of individuals with deep expertise in the focal domain, which are then aggregated to make selection decisions (Azoulay et al. 2019, Bian et al. 2021, Criscuolo et al. 2017, Lane et al. 2021). In this paper, we investigate the effects of dom expertise on the relative importance evaluators place on project novelty and feasibility when evaluating the quality of alternative projects from a firm's R&D pipeline.

Novelty and feasibility represent two distinct dimensions of an idea's quality (Amabile 1996, Audia and Goncalo 2007, Gallo et al. 2018). Whereas novelty relates to the degree a new idea or project departs from the existing knowledge or technological trajectory (Benner and Tushman 2003, Boudreau et al. 2016), feasibility relates to its ease of implementation leveraging existing resources and capabilities (Baer 2012). Although recent work has shed light on the relationship between the novelty of proposed projects and their quality—and has illuminated that evaluation processes tend to discount the value of novel ideas (Boudreau et al. 2016, Criscuolo et al. 2017, Krieger et al. 2021, Wang et al. 2017), what remains relatively less known is how expert evaluators perceive the relationship between the *feasibility* of a proposed project and its *quality*.

The importance of a project's feasibility on evaluation and selection decisions is evident in the technology behind reusable rockets. The idea of reusable rockets is not new and by the laws of physics technically possible, and began with the invention and implementation of rocket technology in the first half of the 20th century (Ross 2018). The effort gained traction after NASA started the Space Shuttle project in the 1960s to create resuable rockets. Yet none of the existing commercial and governmental space enterprises was able to muster the will or consensus to actively invest and/or develop "serious" efforts to make reusable rockets a reality (Ross 2018). This contrasts the tremendous advances made over the past decade through the deliberate strategy of insurgent organizations like SpaceX and Blue Origin (Lag 2022), which have revealed differential assesments about the feasibility of reusable rockets between incumbents and new organizations. The next phase of rocket reuse development will now be slated towards further improvements in its reliability, efficiency and cost (Lag 2022), indicating that uncertainty regarding project feasibility will no longer be a deterrant of continued progress and advancement.

This example shows that innovation progress can be hindered when projects and ideas might appear *infeasible* to individuals with deep expertise in the domain and require people from outside the domain to recognize their value and broad potential for impact. It also suggests that established firms may fail to produce radical innovations not only because novel ideas depart more significantly from their established knowledge and technological trajectories, but also because the ideas, themselves may require different capabilities that do not yet exist in the current marketplace—thereby reducing their feasibility and ease of implementation within a prevailing context. During evaluation processes, both the novelty of a proposed project or idea and its feasibility are likely to be critical drivers of firms' resource allocation decisions.

Despite the importance of novelty *and* feasibility in shaping evaluation processes of quality, there lacks a clear understanding of the systematic effects that novelty and feasibility impart on evaluation and selection decisions. Prior work suggests that domain expertise affects evaluation outcomes by altering how individuals process and weight the relative merits and demerits associated with a new project or idea (Boudreau et al. 2016, Li 2017, Moreau et al. 2001, Pier et al. 2018). While domain expertise may lead evaluators to undervalue novel ideas that depart more substantially from existing technologies and processes (Boudreau et al. 2016, Moreau et al. 2001, Mueller et al. 2012), experts' deeper and more enriched knowledge structures of the domain may improve their accuracy in assessing well-defined and familiar properties of a proposed idea (Chase and Simon 1973), such as its feasibility. Hence, addressing this question of how expertise in the domain affects the relative importance of project novelty and feasibility during evaluation and selection decisions is likely to play a critical role in shaping a firm's technological trajectory.

We collaborated with the U.S. National Aeronautics and Space Administration (NASA) and Freelancer.com to investigate how the novelty, feasibility, and quality of technological solutions impact evaluations. NASA announced a space robotics challenge series and we worked with the agency to recruit our participants using an open "broadcast search" for registrants (Jeppesen and Lakhani 2010). To exogenously vary the evaluators' expertise in the domain, we collected information on all registrants' background and experience, as well as their performance on a skills test in the domain area. Overall, we mobilized a total of 374 unique evaluators that were randomly drawn from both inside and outside the domain area of robotics engineering and exogenously assigned to evaluate 10 out of 101 solutions, for a total of 3,869 evaluator-solution pairs across nine challenges. The evaluators judged each solution according to their novelty, feasibility and quality and provided open-text comments justifying their choices.

We report on several noteworthy patterns. First, we find that experts exhibit a *feasibility preference*, meaning that they rate a solution's quality higher as their judgments of a solution's feasibility increases. Second, we find that experts are more likely to view a solution's novelty and feasibility as tradeoffs in a solution's design, systematically favoring solutions that are high feasibility-low novelty over solutions that are low feasibility-high novelty. We show that these effects are strengthened in the context of high complexity problems that draw on multiple domains and require more effort and skill to solve and evaluate. Third, we leverage text analysis, including word embedding models to analyze the open-text comments to glean insights on how experts and non-experts differ in how they evaluate a solution's novelty and feasibility. We find that experts exhibit deeper level of information processing and attention to the feasibility of a solution but find few meaningful differences between evaluator expertise and a solution's novelty. Moreover, experts are more likely to rely on heuristics during evaluation processes by comparing solutions on the basis of their feasibility and selecting the most feasible alternative, while discounting solutions that are more novel but potentially riskier to implement.

Our findings contribute to the strategy and innovation literature on the role of expertise and evaluations in shaping the direction of technological innovation advances. The insights from this study are likely to have direct implications for innovation managers seeking to greenlight innovative ideas and projects. A direct takeaway of this research is that project novelty and feasibility are two distinct dimensions of a new technological idea. As resource allocation decisions are shaped significantly by prior knowledge overlap to the focal domain, prior experience and expertise can inhibit investment in projects for which their technical feasibility remains unscoped in uncharted territory.

2. Experts and the Evaluation and Selection of R&D Projects

Due to limited resources and time constraints, firms cannot invest in every new idea they generate (Bower 1972, Christensen and Bower 1996, Staw and Ross 1987). The objective of R&D evaluation and selection processes in technical domains is to decide which ideas should be selected for implementation within organizations (Berg 2016, Criscuolo et al. 2017, 2021, Csikszentmihalyi 1999). These strategic decisions, which are often decided by senior members within a firm (Criscuolo et al. 2017), can have critical implications on a firm's knowledge trajectory (Lane et al. 2021), competitive advantage (Moran and Ghoshal 1999), adaptability to changing market and technological conditions (Eisenhardt and Tabrizi 1995, Teece et al. 1997, Tushman and Anderson 1986), and even its survival (Tripsas 1997).

Domain experts are often integral to evaluation and selection decisions. In technical domains, the overseeing authority, often a senior manager within the firm, will rely on the judgments of domain experts with deep knowledge of the discipline or area to assess the potential value of new ideas (Berg 2016, Criscuolo et al. 2017, Lane et al. 2021). The success of these ideas, however, often remains uncertain until after they have been implemented as a product, service or system (Azoulay and Li 2020, Ford 1996, Nelson and Winter 1982, Simonton 1999, Wicht and Szajnfarber 2014). Domain expertise is defined as familiarity with the factual and technical knowledge of the domain, such as the principles, information, opinions and paradigms that individuals develop over time through intensive training and practice—sometimes over multiple years (Dane 2010, Hinds et al. 2001, Shanteau 1992a). As individuals accumulate experience and knowledge performing an activity in a domain or context, their capacity to perform a similar activity in the future tends to improve (Helfat and Peteraf 2015). This is evidenced by a rich line of work on chess research, in which grand masters demonstrate superior memory skills over novices in memorizing chessboard positions due to a storing of intact and well-organized "chunks" of four to five piece chess configurations in their memory (Chase & Simon, 1973; De Groot, 1978). Interestingly, the ability to recall "chunks" from memory disappeared when the chess pieces were instead organized randomly on the chessboard. Similarly, Shanteau (1992) reported that expert auditors are less likely to be influenced by irrelevant information, and research on problem solving in physics suggests experts demonstrate a *deeper* and more enriched understanding of the domain through their categorizing of problems using the major physics principles that would be used to solve the problems, as opposed to the *surface level* entities in the problem statement used by novices (Chi et al. 1981). During evaluation processes, the benefits of domain expertise tend to accumulate as an informational advantage (Henderson 1993, Simon 1955), as closeness to the domain may enable experts to make more *accurate* judgments about the *true* quality of ideas, particularly when they build on existing knowledge and technologies in the domain and allow experts to draw on their prior training and experiences (Li 2017, Shanteau 1992a). On the other hand, other work shows that experts are no better than novices in their forecasting abilities and are also prone to systematic biases in their decision-making (Camerer and Johnson 1991, Hinds et al. 2001, Johnson 1988, Mollick and Nanda 2016, Tetlock 2009). These systematic biases can shape which ideas are selected during evaluations of idea quality (Li 2017, Reitzig and Sorenson 2013).

An important insight is that domain expertise alters how people sample and process information when confronted with a decision task (Camerer and Johnson 1991, Chi et al. 1981, Shanteau 1992a, Simon 1978). During evaluation processes, expertise will likely affect which information cues individuals "sample" (Åstebro and Elhedhli 2006) on the novelty and feasibility of an idea and the relative importance they place on each of these components. The need to get products to market on time in technical domains, such as in new product development, makes the ease of implementing a new idea an essential consideration for driving productivity. Although the objective of idea generation or problem solving is to give rise to novel ideas, originality is not typically a necessary condition for implementation within technical domains (Holmfeld 1970). Rather, the objective is to produce or design a product, process or system that can improve existing technologies and solutions and be put to use in a timely manner (Allen 1977, Florman 2014, Holmfeld 1970, Layton Jr 1974, Woolnough 1991).

This need for execution and timebound deliverables within technical organizations may focus evaluators' attention on the technical feasibility of an idea, as a primary criterion of its quality. This skewed focus may be even greater for domain experts. The feasibility of an idea is an accessible property of a solution that draws most directly on an individual's prior experience and memory of how knowledge "chunks" ought to be connected. Domain experts who are well-versed with existing solutions through their prior experiences with solving similar problems (Florman 2014, Layton Jr 1974, Vincenti 1990), will also be best positioned to make judgments confidently about the feasibility of a proposed solution. In particular, evaluations of an idea's feasibility is most likely to advantage individuals with deep knowledge and understanding of the domain, whose enriched knowledge structures and detailed mental maps of the existing solution space can be best utilized for judging the feasibility of a potential solution (Kornish and Ulrich 2011, Shanteau 1992a, Simon 2019). This superior ability may be exemplified by experts strategically leveraging their knowledge of "broken-leg" cues—rare but highly diagnostic cues that undoubtedly increase the accuracy of predictions (Johnson 1988, Meehl 1954). In the context of evaluations in technical domains, the "broken leg" cue would be considered a "fatal flaw" in a solution design that would make it impossible to develop into a completed product. Critical to "broken-leg" cues are their accessibility only to domain experts, as they enable highly knowledgeable individuals to make confident decisions about a given task or solution. More broadly, domain expertise is most relevant for decisions and tasks that can be broken down by systematic reasoning and analytic processes. Evaluations of a solution's feasibility are similar to procedures like medical diagnosis, chemical reaction paths and other processes that follow "if-then" pair structures (Simon 1987, 2019). These types of decisions take the structure of "if" a set of conditions or patterns are recognized, "then" a body of information associated with the "if" can be evoked from memory and applied to the current situation (Simon 1987). Such structured relationships share a common property of having limited uncertainty in potential outcomes, and enable confidence in one's decisions.

In contrast, a novel idea, by definition, departs from existing solutions in the market (Boudreau et al. 2016). This makes prior knowledge and accumulated experience less applicable to evaluating novel ideas. Experts' mental maps have a tendency to break down when applied to new and unfamiliar terrains. For instance, chess masters perform no better than novices when chess positions are placed randomly on the chessboard, as they are no longer able to draw on well-organized chess configurations in their memory (Chase and Simon 1973, Hardiman et al. 1989), and in creative forecasting, experts appear to have no material advantage over novices when their knowledge needs to be extrapolated into unfamiliar areas with unknown solutions (Kornish and Ulrich 2014, Mollick and Nanda 2016). Moreover, experts have difficulty putting aside their past experience when predicting how novices will approach problems or respond to events (Hinds 1999, Hinds et al. 2001), and exhibit inflexibility to accommodating new rules and principles (Camerer and Johnson 1991, Luchins 1942). Perhaps one of the most well-known downsides of expertise is from studies on problem solving and *functional fixedness*, which occurs when people struggle to solve problems using unfamiliar and novel approaches after being previously exposed to a familiar approach (Adamson 1952, Duncker 1945). Such mental blocks are likely to hinder domain experts from seeing the potential upside of novel ideas during evaluation—leading them to discount their value.

Along with being atypical compared to existing solutions, novel ideas also possess greater risk, which can be described as the extent to which there is uncertainty about whether a potentially significant or disappointing "extreme" outcome of a decision will be realized (Sitkin and Pablo 1992, Sitkin and Weingart 1995). Risk increases variability in the outcome distribution, raising the uncertainty to which a given outcome can be predicted (Sitkin and Pablo 1992). When judging technical solutions, this risk typically cannot be completely resolved without experimentation, prototyping and iterating (Cannon and Edmondson 2005, Franzoni et al. 2021, Thomke 1998). Consequently, there is limited information available about a novel idea (MacCrimmon et al. 1988) from which its potential outcomes can be assessed in advance on the basis of existing evidence (Fox and Tversky 1995).

Novel ideas may appear even riskier to experts because they can challenge experts' competence and knowledge in the domain. Several experimental studies suggest that people experience ambiguity aversion when they feel incompetent in a domain (Fox and Tversky 1995, Franzoni and Stephan 2021). Ambiguity aversion is heightened when people need to compare two tasks, one for which they have superior knowledge about and the other in which they have limited knowledge. The differences in states of knowledge between the two tasks undermines their confidence in the unfamiliar task (Fox and Tversky 1995). Following this logic, when asked to evaluate the novelty and feasibility of an idea, experts may begin to compare and contrast the differences in their states of knowledge across the two components. This may create a preference for focusing on what they know and can control—i.e., evaluating feasible ideas, and an aversion towards what they know little about—i.e., evaluating novel ideas.

Within technical domains, the risks associated with novel ideas may create an even greater aversion to them among experts. In particular, failure to adequately account for the risks associated with a particular decision can lead to catastrophic consequences that may jeopardize a firm's legitimacy (Chai et al. 2021). This can shape preferences that value certainty (or risk avoidance) more heavily than uncertainty (Douglas and Wildavsky 1983), creating asymmetric risk-reward payoffs. This contrasts the view of novelty as high-risk high-reward in domains, such as the creative arts, science and venture capital investing, where a "hit" may lead to extraordinary attention, success and recognition (Baum and Bird 2010, Berg 2022, Uzzi et al. 2013). In contrast, in technical domains, where there is a fundamental need to produce a working design, product, or system (Vincenti 1990), the dire consequences of failure are likely to outweigh the benefits of creating a novel design for which its reliability and performance cannot be predicted ex ante. An understanding of these asymmetric risk-reward payoffs in technical domains are likely to weigh more on experts, who are deeply embedded in the culture and expectations of the domain, as well as the negative repercussions of failure associated with high-risk and uncertain ideas. Consequently, we might expect that domain experts are more likely to demonstrate conservative judgments, skewing their judgments in favor of less novel solutions with lower outcome uncertainty.

In short, technical organizations have a need to put solutions to use (Layton Jr 1974, Vincenti 1990, Woolnough 1991), which makes implementation an important goal of evaluation processes. Hence, we might expect evaluators to overvalue the feasibility of an idea over its novelty. This greater attentional focus on solution feasibility may be even higher for domain experts due to their deep understanding of existing solutions and direct applicability of their prior knowledge in judging the feasibility of an idea. Following this logic, domain experts may make evaluations of quality by first anchoring on the feasibility of an idea and then making adjustments for an idea's novelty, after making assessments about its feasibility (Kahneman et al. 1982, Tversky and Kahneman 1974). Due to a strong desire to mitigate risk and reduce uncertainty in outcomes, we might expect experts to exhibit a feasibility preference, valuing solutions that are higher in feasibility but lesser in novelty, resulting in a tradeoff between idea novelty and feasibility during evaluations of quality.

Hypothesis 1 (Feasibility Preference). Domain experts are more likely to value technical solutions that are higher in feasibility but lower in novelty during evaluation processes of quality.

Problem complexity is a key feature of the innovation process, and it shapes evaluation outcomes in important ways. In R&D problems, complexity refers to the number of technological domains represented in the problem (Boudreau et al. 2011). Higher complexity problems are more demanding on one's attention, time and effort, because they require an understanding of knowledge from different domains and a need to draw upon different knowledge areas in order to be solved effectively (Fleming 2001, Kavadias and Sommer 2009). Drawing on Simon (1969), who conceptualized problem solving as a "search" through multiple solutions, complex problems offer multiple ways to search and build upon existing solutions. These alternative approaches may be interdisciplinary in nature, while also yielding solutions that may have different levels of feasibility (Boudreau et al. 2011, Jeppesen and Lakhani 2010). Whereas a low complexity problem may only require solution approaches from one discipline or field, those that draw on multiple domains are more complex and often necessitate that solvers *recombine* knowledge and solutions from different technological fields (Ferguson and Carnabuci 2017, Rosenkopf and Nerkar 2001). Hence, higher problem complexity is also likely to be associated with a greater share of less feasible, low performing ideas—particularly as it becomes more difficult to solve new problems with past approaches (Kornish and Ulrich 2011, Sommer et al. 2020).

Complexity can lead to greater recombinant uncertainty as the number of different technological components increases (Fleming 2001, Nelson and Winter 1982, Taylor and Greve 2006, Weitzman 1998). Although each component may be feasible on its own, each component adds an additional layer of uncertainty about the entire system's ability to satisfy the necessary solution requirements. Due to the need to recombine across technological boundaries—which increases the number of interconnections—solutions to highly complex problems insert "randomness" into the system architecture (Hennig et al. 2022) that can arouse greater uncertainty and risk in terms of their potential for implementation (Fleming 2001, Kaplan and Vakili 2015). Following this logic, experts may be even more likely to turn their attention predominantly to the feasibility of a solution to a highly complex problem—focusing on whether a solution can be implemented as their primary criterion of quality. We hypothesize that experts will be even more likely to prioritize a solution's feasibility over its novelty for higher complexity problems.

Hypothesis 2 (Problem complexity). The tradeoff between the feasibility and novelty of technical solutions during expert evaluations of quality is larger when the R&D problem has higher complexity.

3. Research Design

In this section, we describe the setting and research design, and provide details on the evaluator recruitment, procedures, random assignment, and key measures.

3.1. Setting and Recruitment of Evaluators

We carried out our research in the context of an evaluation process for technical solutions to R&D problems that were part of NASA's Abstrobee Challenge Series. Astrobee is NASA's free-flying robotic system, which is designed to complete routine tasks such as taking inventory, documenting experiments conducted by astronauts and moving cargo throughout the station, freeing up time for astronauts to focus on activities only humans are capable of doing. Astrobee is an integral part of NASA's mission to return to the Moon as well as other deep space missions.

In 2018, together with Freelancer.com, a freelance marketplace website that allows potential employers to post jobs that freelancers can then bid to complete, NASA and a team of researchers launched the Astrobee Challenges Series, leveraging the freelancer community for solutions for an attachment and orientation arm (see Szajnfarber et al. 2020). NASA launched a total of seventeen "challenges", with a total prize money of $25,000, with individual prizes ranging from $250 to $5,000. Each challenge asked for solutions that varied from a particular piece of the attachment arm to the entire arm design, with the objective being that the winning solutions would be incorporated into future robotic arm designs to be used with Astrobee. Across the seventeen contests, more than 250 solutions were submitted.

After the contests were completed, NASA offered to purchase any solutions submitted to the original challenge series to be used in future research. 80% of problem solvers responded. In May 2021, we partnered with Freelancer.com to recruit freelancers to help with solution evaluation for a subset of nine (of the 17) challenges and all of the 101 purchased solutions within the nine challenges. We selected these nine challenges to create heterogeneity in challenge (or problem) complexity and solution quality within each challenge. We broadcasted to registrants that the purpose of the evaluation effort was to help NASA understand how the community can potentially assist in evaluating solutions to engineering challenges for NASA and other organizations. In addition, we told all potential registrants that the task would consist of evaluating 10 original solutions from two challenges (five solutions from each challenge) for their novelty, feasibility and overall quality. We communicated that the entire evaluation process would take an estimated time of 60-90 minutes, and that we would pay each evaluator $25 upon completion of their evaluation task. The opportunity was advertised on the Freelancer.com website and attracted 18,765 registrants to the call. Each registrant completed an initial survey which included an human resources (HR) screen on their demographic information (e.g., Freelancer.com user name, gender, age, country, educational background, work experience in a technical organization outside of their educational experience, expertise in robotics and related disciplines) and a skills assessment that included 17 technical questions from the domain area that we pre-tested on individuals with different levels of expertise in the domain of robotics and related disciplines.

3.2. Evaluator Selection and Evaluation Procedures

Given our interest in generating variation in evaluator expertise in the domain, as well as replication and increased degrees of freedom, we selected roughly equal groups of evaluators from three distinct groups among the 18,765 registrants: (i) those from the *unscreened* pool of registrants, (ii) those who passed the *Skills test screen* threshold of 13 or more out of 17 (>75%) on the skills test, (iii) those who passed the *HR screen* threshold of two or more years of work experience in the domain of robotics/mechatronics engineering. We used the *Skills test screen* and *HR screen* as two alternative approaches for screening individuals for expertise in the domain area, namely high proficiency, and prior disciplinary training and experience. We generated these groups by first randomly selecting individuals from the unscreened pool of registrants and inviting them to participate. After removing these selected individuals, we then rank ordered individuals by their skills test score, and invited everyone who scored a 13 or more out of 17 to participate. Finally, after removing the individuals selected for the unscreened and skills test screen evaluator groups, we rank ordered all remaining registrants by their number of years of work experience in robotics or mechatronics engineering and selected every registrant who had two or more years of work experience to participate. 549 evaluators accepted our invitation to participate, and 374 completed the evaluation task. This generated roughly equal numbers of evaluators from each of the three groups (125 from the general pool, 109 from the HR screen, and 140 from the skills test screen).

Overall, our assignment of evaluators to solutions created 3,869 evaluator-solution pairs. We used a randomized block design where we first, randomly assigned each evaluator two of the nine challenges to evaluate, and then randomly assigned them five solutions from each challenge to evaluate, for a total of 10 solutions per evaluator. The random assignment of evaluators to challenges and solutions was critical to the experimental design because it created exogenous variation in the solutions that each evaluator was exposed to, while enabling them to compare across solutions within each challenge they were assigned to review. In Table A1, we show that the randomized block design achieved balance across the evaluator covariates.

For each challenge, evaluators were given a general overview of the challenge, and then asked to download and read the original problem statement as well as familiarize themselves with the submission guidelines associated with the problem. After reading and familiarizing themselves with the problem statement and submission guidelines, each evaluator then proceeded to download the solution (a pdf with roughly 10-15 pages of designs and explanation), provided each solution with both numerical ratings and narrative comments for its feasibility, novelty, and overall quality, as well as report their confidence for each rating (see Figure A1 for the first page from a sample solution document). The narrative comments consisted of open-text responses where the evaluators were asked to document all the factors or aspects that led to their rating (see Figure A2 for screenshots of evaluation procedures).

Figure 1 shows a conceptual flow of the evaluation process, randomization of evaluators to challenges and solutions, as well as the evaluation procedures. Table 1 shows the number of ratings by challenge by evaluator expertise as well as the overall mean number of evaluations per solution for each challenge.

[Figure 1 about here]

[Table 1 about here]

3.3. Dependent Variables

Our main dependent variable, *Quality rating,* is measured on a Likert scale from 1 to 7 and corresponds to the evaluator's quality rating of a solution.

3.4. Independent Variables

3.4.1. Evaluator Expertise Type

We use the categorical variable, *Evaluator expertise type* to differentiate between the three types of expertise: *Unscreened*, *HR screen*, and *Skills test screen.*

In Table 2, we compare the three groups of evaluators on four dimensions of expertise to obtain a better understanding of the evaluators' level of knowledge and experience in the focal domain of robotics. The first dimension, years of robotics work experience, is the criterion used for the HR screen. Here, we observe that while the HR screen evaluators have significantly more years of work experience than the unscreened and skills test screen evaluator groups (F(2,369) = 63.72, $p < 0.01$), both the unscreened and skills test screen evaluator groups have on average, less than one year of work experience in the focal domain area of the evaluation task. Figure A3 shows the distribution of robotics work experience by the three types of evaluator expertise. The second dimension, skills test score, is the criterion used for the skills test screen. Here, we observe that the skills test score increases from the unscreened evaluators (mean = 7.128, s.d. = 3.245) to the HR screen evaluators (mean = 9.62, s.d. = 3.188) to the skills test screen evaluators (mean = 13.856, s.d. = 1.060) out of a total possible score of 17, and the means are significantly different (F(2,369) = 221.80, $p < 0.01$). However, it is worthwhile to note that on average, even the unscreened evaluators were able to score over 50 percent on the skills test. Given that the skills test was multiple choice with five or six possible responses, we would expect an evaluator to score closer to 15-20 percent if they were to guess the correct response to each question. Figure A4 shows the distribution of skills test scores by the evaluator expertise type.

We also report two alternative dimensions of evaluator expertise in the domain that were collected as part of the registration survey. The first is Jeppesen and Lakhani (2010)'s measure of expertise distance and captures the self-perceived distance between the evaluators' own field of expertise and the focal domain of the solutions. The variable is based on the answer to the survey question: "A robotics design problem is: 1—inside my field of expertise, 3—at the boundary of my field of expertise, 5—outside my field of expertise." Respondents chose any value between 1 and 5 on a Likert scale. The higher the score, the greater the perceived distance to the field of robotics. As shown in Table 2, whereas both the HR screen (mean = 1.917, s.d. = 1.033) and skills test screen (mean = 2.267, s.d. = 1.053) indicated that a robotics design problem is inside their field of expertise, the unscreened evaluators indicated that a robotics design problem is on average closer to the boundary of their field of expertise (mean = 3.528, s.d. = 1.411). Figure A5 shows the distribution of *Robotics design expertise distance* by evaluator expertise type.

Our final measure uses Szjanfarber et al. (2020)'s similarity to a roboticist scale, which conceptualizes the disciplinary distance between an evaluator's prior experience and training and the focal domain of the solutions as either 1—within discipline, 2—relevant engineering experience, 3—science and technology experience, and 4—other. Respondents indicated the number of years of work experience they had accumulated in each discipline, and the measure was coded based on their responses. As shown in Table 2, the unscreened evaluator group on average spans the boundary between relevant engineering experience (e.g., work experience in electrical engineering, software or computer science or engineering drawing), and science and technology (e.g., design, project management, aerospace and defense, architecture or design). By construction, the HR screen evaluators are within discipline (because they were selected based on having prior work experience in robotics) and the skills test screen evaluators tend to have relevant engineering experience, with some of them being within discipline. Figure A6 illustrates the *Distance to roboticist discipline* by evaluator expertise type.

Taken together, these measures of evaluator expertise in the domain indicate that although screening the registrants—either by the HR screen or skills test scree—created meaningful variation in the evaluators' distance to the focal domain in robotics, all three evaluator groups were drawn from a pool of registrants who had accumulated prior training and work experience in technical disciplines in either engineering or science and technology more broadly. Importantly, we can draw a boundary on our evaluator pool as having at least some basic understanding and knowledge of robotics design problems.

[Table 2 about here]

3.4.2. Novelty Rating

We use the variable, *Novelty rating*, measured on a Likert scale from 1 to 7, to denote the evaluator's novelty rating of a solution.

3.4.3. Feasibility Rating

We use the variable, *Feasibility rating*, measured on a Likert scale from 1 to 7, to denote the evaluator's feasibility rating of a solution.

3.4.4. Feasibility Preference

We measure the extent that evaluators focus more extensively on the feasibility of a solution over its novelty using the variable *Feasibility Preference*, which corresponds to the difference between the *Feasibility rating* and the *Novelty rating* given by an evaluator to a solution, as follows:

$$Feasibility\ preference = Feasibility\ rating - Novelty\ rating \quad (1)$$

We also model *Feasibility preference* as a categorical variable, taking three possiblie values: *Novelty rating* > *Feasibility rating*, *Novelty rating* = *Feasibility rating*, and *Feasibility rating* > *Novelty rating.* We show in the results section (see section 4.1) that our econometric results are robust to both specifications.

3.4.5. High Problem Complexity

We use the binary variable, *High Problem Complexity,* which took a value of 1 if a problem was a high complexity problem and 0 otherwise. We use the prize money (in US dollars) awarded to the winning solution as a proxy of the problem complexity. The prize money for each solution was independently determined by two external domain experts (i.e., external to NASA; see Szajnfarber et al., 2020), and was based on an estimate of the number of domains represented by the problem, as well as the amount of time and effort required to come up with a solution. Figure 2 plots the prize money corresponding to each of the nine challenges. Based on the prize money, we categorize three challenges, SAM (Smart Attachment Mechanism), SPAM (Smart Positioning and Attachment Mechanism) and SRA (Smart Robotic Arm), with a prize money of $1500 or more as being high problem complexity and the remaining six challenges, SDM (Simple Deployment Mechanism), MDC (Mechanically Driven Clamp), PSA (Positioning Software Architecture), HMSA (Health Monitoring Software Architecture), RASA (Robotic Arm Software Architecture) and MIS (Material Interface Surface) as being low problem complexity.

3.4.6. Other Variables

Our analysis relies most heavily on the research design's randomization and exploitation of multiple observations per solution, with a series of dummy variables for each unique solution. Since prior work has noted that confidence varies with a decision-maker's expertise in the domain (Kahneman and Klein 2009, Tversky and Kahneman 1974), we control for the evaluator's confidence in their novelty and feasibility ratings using *Novelty confidence* and *Feasibility confidence*, respectively. We also control for several evaluator demographic characteristics (gender, age range, level of education, U.S. citizen/residing in U.S.), which have been shown to affect decision-making (Weber and Johnson 2009).

3.5. Econometric Approach

We use ordinary least squares (OLS) models to estimate the how quality is related to a solution's novelty, feasibility by evaluator expertise type. First, we investigate how *Evaluator expertise type* shapes the relative weighting of novelty and feasibility in solution quality using the following OLS model in (2):

$$Quality\ rating_{ij} = \beta_0 + \beta_1 Evaluator\ expertise\ type_i + \beta_2 Novelty\ rating_{ij} +$$
$$\beta_3 Feasibility\ rating_{ij} + \beta_4 Evaluator\ expertise\ type_i \cdot Novelty\ rating_{ij} +$$
$$\beta_5 Evaluator\ expertise\ type_i \cdot Feasibility\ rating_{ij} + \beta_6 Novelty\ confidence_{ij} +$$
$$\beta_7 Feasibility\ confidence_{ij} + \beta_8 X_i + \gamma_j + \varepsilon_{ij}, \qquad (2)$$

where we control for the evaluator i's confidence in their novelty and feasibility rating for solution j, evaluator covariates X_i, and solution fixed effects, γ_j. The solution fixed effects allow us to make comparisons between evaluators randomly assigned to evaluate the same solution facilitating within solution differences by *Evaluator expertise type*.

Second, we examine the extent that evaluators exhibit a *Feasibility preference* when rating a solution for its quality using the OLS model in (3):

$$Quality\ rating_{ij} = \beta_0 + \beta_1 Evaluator\ expertise\ type_i + \beta_2 Feasibility\ preference_{ij} + \beta_3 Evaluator\ expertise\ type_i \cdot Feasibility\ preference_{ij} + \beta_4 Novelty\ confidence_{ij} + \beta_5 Feasibility\ confidence_{ij} + \beta_5 X_i + \gamma_j + \varepsilon_{ij}. \quad (3)$$

4. Results

4.1. Econometric Results

In Table 3, we provide summary statistics on the evaluators by *Evaluator expertise type*. We observe that overall, our evaluator pool is highly male, between 25-34 years old on average, highly educated and mostly residing outside the US (by country, India has the highest share of evaluators with 23%, with the remaining 72 countries holding between 0-6% of the share).There is some variation across evaluator groups, with the unscreened evaluators more likely to be female, younger in age, and less likely to hold a Bachelor's or Master's degree than the screened evaluator groups. Also, the HR screen evaluators are more likely to be from the US. Table 4 provides the correlation table of the main variables used in the analyses.

[Table 3 about here]

[Table 4 about here]

4.1.1. Results on Feasibility Preference (Hypothesis H1)

Hypothesis 1 theorized that domain experts are more likely to rate the quality of a solution higher when the solution design is higher in feasibility and lower in novelty, corresponding to a tradeoff between the feasibility and novelty of a solution design. To test Hypothesis 1, first, in Table 5, we use OLS models to examine the relationship between *Evaluator expertise type* and the component ratings of *Novelty* and

Feasibility on the *Quality rating* of the solution. Then in Table 6, we examine the relationship between *Evaluator expertise type*, *Feasibility preference* (corresponding to *Feasibility rating* - *Novelty rating*; see section 3.4.4 for details) and the *Quality rating of the solution*.

Turning first to Table 5, Model 1 regresses *Quality rating* on the *Evaluator expertise type*, Model 2 adds the *Novelty* and *Feasibility ratings*, Model 3 adds the interaction term between *Evaluator expertise type* x *Novelty rating*, Model 4 adds the interaction between *Evaluator expertise type* x *Feasibility rating*, and Model 5 adds both interaction terms. Then, Models 6-9 add confidence ratings for novelty and feasibility (Model 6), evaluator covariates (Model 7), as well as challenge (Model 8) and solution (Model 9) dummies.

In Model 1, we observe that domain expertise leads to more critical scores. Compared to *Non-experts*, both the *HR screen* (Model 1: -0.353, $p < 0.01$) and *Skills test screen* (Model 1: -0.576, $p < 0.01$) evaluators gave lower scores on average. These results are consistent with prior work on evaluation scores and domain expertise in the subject area, which have shown that evaluators who are closer to the domain area give more critical evaluation scores (Amabile 1983, Boudreau et al. 2016). Model 2 adds the novelty and feasibility ratings, and we observe that both have a positive relationship with the quality of the solution (Novelty rating: 0.362, $p < 0.01$; Feasibility rating: 0.578: $p < 0.01$). Next, in Model 3, we observe that the coefficients for the interaction term between *Evaluator expertise type* x *Novelty rating* is positive for the *HR screen* (Model 3: 0.0275, *ns*) and positive and significant among the *Skills test screen* evaluators (Model 3: 0.0621, $p < 0.01$). Turning to Model 4, we observe that the coefficients for the interaction term between *Evaluator expertise type* x *Feasibility rating* is positive and significant for both the *HR screen* (Model 4: 0.101, $p < 0.01$) and *Skills test screen* (Model 4: 0.131, $p < 0.01$) evaluators. Model 5 includes interaction term between *Evaluator expertise type* and both novelty and feasibility, and we observe that while the coefficients for the interaction term are positive and significant with feasibility (HR screen: 0.141, $p < 0.01$; Skills test screen: 0.151, $p < 0.01$), neither coefficient for the interaction term is significant with novelty (HR screen: -0.0611, *ns*; Skills screen: -0.0313, *ns*). This suggests that domain experts place greater weight on the feasibility of a solution when evaluating the quality of technical ideas, compared non-experts.

Moreover, these relationships remain robust after adding the confidence ratings, evaluator covariates, challenge and solution dummies in Models 6-9. Figures A7 and A8 illustrate the margins plot with 95% confidence intervals (CIs) estimating the relationships between the solution's *Quality rating*, *Evaluator expertise type*, novelty (Figure A7) and feasibility (Figure A8) ratings from Table 5 Model 5; the figures show that whereas there are minimal differences in how evaluator expertise type affects the relationship between a solution's novelty and its quality, expert evaluators (i.e., both HR screen and Skills test screen groups) are more likely to prioritize a solution's feasibility as a critical predictor of the solution's quality.

[Table 5 about here]

Next, in Table 6, we examine the relationships between *Evaluator expertise type*, *Feasibility preference* and a solution's *Quality rating.* In Model 1, we add *Feasibility preference*, Model 2 adds the interaction term between *Evaluator expertise type* and *Feasibility preference*, Model 3 adds the confidence ratings for novelty and feasibility, and Model 4 adds the evaluator covariates. Finally, Models 5 and 6 add challenge and solution dummies, which allows us to examine differences within challenge, and within solution, respectively.

Turning to Model 1, we observe that the *Feasibility preference* has a positive relationship with *Quality rating* (Model 1: 0.104, $p < 0.01$). This suggests that evaluators rate solutions as higher in quality when they are higher in feasibility and lower in novelty (recall that *Feasibility preference* is the difference between the solution's *Feasibility rating* and *Novelty rating*). Next, Model 2 adds the interaction term between *Evaluator expertise type* x *Feasibility preference.* We observe that the coefficients for *HR screen* x *Feasibility preference* (Model 3: 0.133, $p < 0.01$) and *Skills test screen* x *Feasibility preference* (Model 3: 0.148, $p < 0.01$) are both positive and significant. This suggests that experts are more likely than non-experts to prefer solutions that are more feasible but less novel.

To gain deeper insights into these relationships, Figure 3 shows the margins plot with 95% confidence intervals (CIs) between *Quality rating*, *Evaluator expertise type*, and *Feasibility preference.* Here, we observe that both the *HR screen* and *Skills test screen* evaluators view feasible solutions to be higher in quality, and that novelty and feasibility are viewed as tradeoffs in a solution's design.

Interestingly, there is no evidence of a tradeoff between novelty and feasibility among the unscreened evaluators. Moreover, the observed relationships remain stable and robust when we control for the evaluator's confidence ratings in Model 3, evaluator covariates in Model 4, as well as challenge and solution fixed effects in Models 5 and 6. We note that Models 5 and 6 correspond to our most stringent comparisons, as they examine within challenge and within solution differences, respectively. As robustness, we perform the OLS regression analyses examining the relationship between *Quality rating*, *Evaluator expertise type*, and a categorical variable for *Feasibility preference* (see section 3.4.4). The results remain robust to the alternative specification of *Feasibility preference* in Table A2 and Figure A9.

[Table 6 about here]

[Figure 3 about here]

Taken altogether, the results in Tables 5 and 6 indicate that domain experts prefer solutions that are more feasible but less novel compared to solutions that are more novel but less feasible—perceiving the two components as tradeoffs in a solution's design. Hence, we find support for Hypothesis 1.

4.1.2. Results on Feasibility Preference and Problem Complexity (Hypothesis H2)

Next, Hypothesis 2 theorized that the feasibility preference among domain experts would be larger for more complex problems. First, we present the OLS regression results estimating the relationships between *Quality rating*, *Evaluator expertise type* and the *Feasibility preference* for low complexity (Table 7) and high complexity problems (Table 8). In both Tables 7-8, we begin by modeling the relationship between *Quality rating* and *Evaluator Expertise Type* in Model 1. We then add *Feasibility preference* in Model 2, and the interaction term between *Evaluator Expertise Type* x *Feasibility preference* in Model 3. Models 4-5 add the confidence ratings and evaluator covariates, and Models 6-7 add challenge and solution fixed effects, respectively.

Turning to Table 7, which examines low complexity problems, in Model 1, we observe that domain experts give more critical scores in both the *HR screen* (Model 1: -0.390, $p < 0.01$) and *Skills test screen* (Model 1: -0.614, $p < 0.01$) conditions. We also observe in Model 2 that *Feasibility preference* has a positive and significant relationship with *Quality rating* (Model 2: 0.0900, $p < 0.01$). Turning to the interaction term between *Evaluator expertise type* x *Feasibility preference* in Model 3, we observe that although the coefficients for the interaction term are positive, the coefficients are not significant for the *HR screen* (Model 3: 0.0866, *ns*) or *Skills test screen* (Model 3: 0.0488, *ns*) evaluator groups, indicating that there is no evidence that domain expertise affects the relationship between perceptions of novelty and feasibility being tradeoffs among low complexity problems. These results are consistent in Models 4-7, which add the confidence controls and evaluator covariates, followed by the challenge and solution fixed effects.

[Table 7 about here]

Next, we present the results for high complexity problems in Table 8. Consistent with low complexity problems, in Model 1, we observe that domain expertise leads to more critical evaluation scores for the *HR screen* (Model 1: -0.299, $p < 0.05$) and *Skills test screen* (Model 1: -0.505, $p < 0.01$) evaluator groups. Also, the relationship between *Feasibility preference* and *Quality rating* is positive and significant in Model 2 (Model 2: 0.136, $p < 0.01$). In Model 3, we add the interaction term between *Evaluator expertise type* x *Feasibility preference* and observe that the coefficients are positive and significant for both *HR screen* x *Feasibility preference* (Model 3: 0.225, $p < 0.01$) and *Skills test screen* x *Feasibility preference* (Model 3: 0.361, $p < 0.01$). The reported coefficients remain positive and significant after we add the confidence ratings in Model 4 and evaluator covariates in Model 5. In Models 6 and 7, which include the challenge and solution fixed effects, respectively, the coefficient for *HR screen* x *Feasibility preference* remains positive (Model 6: 0.170, *ns*; Model 7: 0.119, *ns*) and is positive and significant for *Skills test screen* x *Fesibility preference* (Model 6: 0.339, $p < 0.10$; Model 7: 0.276, $p < 0.01$).

[Table 8 about here]

The results in Tables 7 and 8 suggest that feasibility preference among domain experts is larger for high complexity problems, particularly among the *Skills test screen* evaluator group. In supplementary analyses, we show that the results remain robust under alternative specifications that examine the three-way interaction between *Feasibility preference*, *Evaluator expertise type*, and *High complexity* (Table A3).

Overall, our results partially support Hypothesis 2, indicating that the tradeoff between solution feasibility and novelty is larger for high complexity problems among domain experts than non-experts in the *Skills test screen* evaluator group.

4.2. Text Analysis: Insights into the Feasibility Preference

Next, we turn to the open-text comments to gain deeper insights into the econometric results which showed that experts exhibited larger a feasibility preference, being more likely to view novelty-feasibility tradeoff in their evaluation scores than non-experts. First, we leverage Linguistic Inquiry Word Count (LIWC) to content code the open-text comments to examine differences in attentional focus to novelty and feasibility by evaluator expertise type. Prior work suggests that evaluation processes demand a significant amount of attention (Criscuolo et al. 2017, 2021), which suggest that evaluators may decide to prioritize some criteria over others. Second, we leverage word embedding models, a type of unsupervised machine learning approach, to better understand the decision strategies used by the evaluators to come up with their evaluation scores.

4.2.1. Attentional Focus and Depth of Information Processing

We leverage LIWC to measure differences in how evaluators allocated their attentional resources between judging the novelty and feasibility of solutions. LIWC is a text analysis program that counts words in psychologically meaningful categories, and is has been validated across experimental settings for capturing attentional focus and complexity of information processing (Pennebaker et al. 2015, Tausczik and Pennebaker 2010).

We use four related measures to capture differences in attentional focus and depth of information processing in the evaluators' novelty and feasibility comments accompanying each solution. First, we measure the number of words each evaluator used in their open-text comments for novelty and feasibility using the *word count* category in LIWC. Text length or word count is one commonly used measure of idea quantity or depth (Blumenstock 2008, Dimitriadis and Koning 2022). Second, we measure the proportion of six letter words used in each comment leveraging the LIWC category *sixltr*. The proportion of six letter words is often used as an indicator of cognitive complexity, which is associated with deeper thinking and higher levels of reasoning (Pennebaker et al. 2003, Tausczik and Pennebaker 2010). Third, we examine the average number of words per sentence using the *WPS* category in LIWC. Prior work has also established that the sentence length is an indicator for complexity and detail in language (Dimitriadis and Koning 2022, Tausczik and Pennebaker 2010). Fourth, we use the LIWC category *Analytic*, which measures the degree to which an evaluator uses words that suggest formal, logical, and hierarchical thinking patterns as opposed to language that is more intuitive and personal (Jordan et al. 2019).

4.2.2. LIWC Results on Attentional Focus by Evaluator Expertise

In Table 9, we present the means and standard deviations of the attentional focus measures by *Evaluator expertise type* on the evaluators' novelty and feasibility comments. We use one-way ANOVA and Tukey's honestly significant difference (HSD) post-hoc tests to detect differences in the data. Turning to the novelty comments, we observe a few significant differences by evaluator expertise. First, we note that there is a small difference in average comment length (word count) between the *HR screen* and *Unscreened* evaluators ($p < 0.01$), with the *Unscreened* evaluators writing more than the *HR screen* evaluators on average, which is contrasted with a higher proportion of six letter words being used in the *HR screen* ($p < 0.05$) and *Skills test screen* ($p < 0.01$) evaluators compared to the *Unscreened* evaluators—potentially indicating higher levels of thinking.

Turning to the feasibility comments, the patterns suggest that the *Skills test screen* evaluators wrote longer comments ($p < 0.01$), used a greater proportion of six letter words ($p < 0.01$) and more analytic words ($p < 0.01$) in their comments compared to the *Unscreened* evaluator group. We see some evidence that the *HR screen* evaluators engaged in deeper information processing than unscreened evaluators, using a higher proportion of six letter words ($p < 0.01$) and more analytic words ($p < 0.05$) than the *Unscreened* evaluators, but there was no significant difference in comment length (word count) and their comments contained fewer words per sentence than the unscreened evaluators ($p < 0.01$).

Overall, these results are consistent with the notion that domain experts, particularly in the *Skills test screen* evaluator group, are more likely to focus their attention on the feasibility of the solution, showing deeper information processing of a solution's feasibility compared to non-experts. In contrast, we find few significant differences by evaluator expertise type for the novelty of a solution. This suggests that domain experts were better at diagnosing and processing information about a solution's feasibility— potentially "by applying one's 'professional judgment" (Simon, 1987, p. 59) to evaluate a solution's feasibility rapidly and intuitively.

[Table 9 about here]

4.2.3. Decision Strategies By Evaluator Expertise Type

Building on the psychology literature on judgment and decision-making (JDM) (e.g., Camerer & Johnson, 1991; Einhorn, 1974; Gigone & Hastie, 1997; Kahneman et al., 1982), we examine the evaluators' open-text comments for novelty and feasibility to identify common decision strategies used by the evaluators to arrive at their evaluation scores and the extent that use of specific decision strategies differed by *Evaluator expertise type*. We draw on Goldstein and Weber (1995) as well as Rettinger and Hastie (2001)'s approach of content coding decision strategies from open-ended text. The possible strategies were:

1. *Choose the best*. "I found the best solution."
2. *Best compared to others.* "This is a better/worse solution compared to others."
3. *Avoid the worst.* "I found the worst solution and avoided choosing it."
4. *Avoid risky solutions.* "I found the solution that would be the least risky."

Both *"Choose the best"* and *"Best compared to others"* are examples of decision strategies that might be used to select high quality solutions (Gigerenzer and Goldstein 1996, Schwartz et al. 2002, Simon 1957, Weber and Johnson 2009). In contrast, *"Avoid the worst"* and *"Avoid risky solutions"* correspond to two alternative decision strategies for avoiding disappointing outcomes that are low in quality (Sitkin and Pablo 1992, Sitkin and Weingart 1995, Tversky and Kahneman 1974). Both decision strategies demonstrate risk-averse behaviors due to their focus on stability and certainty of outcomes.

We leverage word embedding models to determine whether the open-text comments contained evidence of the decision strategies being used. First, we identified suitable target words corresponding to each strategy, and then used word embedding models to find synonyms of the target word in the evaluators' open-text comments corresponding to their novelty and feasibility ratings for a given solution design. Word embeddings are a type of word representation where individual words are represented as real-valued vectors in a predefined vector space (Bengio et al. 2003). Words with similar semantic meanings and usage have a similar representation in vector space (Bengio et al. 2003). We used a pretrained embedding model called Global Vector or GloVe—an unsupervised learning algorithm for obtaining vector representations trained on the nonzero entries of a global word-word co-occurrence matrix from a 6 billion word corpus—to identify comments containing synonyms associated with each target word of interest (Pennington et al. 2014).

After pre-processing the comments (i.e., removing duplicates, punctuation, stop words, non-English responses), we computed the cosine similarity between each target word vector and each word in the comment to identify potential synonyms with a threshold cosine similarity measure of 0.6 or higher. In Table 10, we show the target word selected for each decision strategy as well as the target word synonyms that were identified from the evaluator comments.

For example, for the decision strategy, *"Choose the best"*, Table 10 shows that we used the target word vector, *"Best"*, and applied our algorithm to identify synonyms associated with *"Best*" from the evaluators' comments. Based on a 0.6 threshold, there were seven synonyms of *"Best"* that were flagged in the comments: *"Best"*, *"Ever"*, *"Excellent*", *"First"*, *"Performance"*, *"Success"* and *"Winning"*. For each decision strategy, we created a dummy variable that took a value of 1 if the comment contained at least one target word synonym, and 0 otherwise. The dummy variables corresponding to each decision strategy were *Novelty_best*, *Novelty_compared*, *Novelty_worst*, *Novelty_risky*, *Feasibility_best*, *Feasibility_compared*, *Feasibility_worst*, and *Feasibility_risky.* As an example of how this algorithm was applied to the open-text responses, consider the following sample comment:

> *"Based on mass of the ESRA its look superior in comparison with SRA-3. However, its operation is not justified under worst environmental constraints."*

After pre-processing the comment, we are left with the following words from the comment:

> *['Based', 'mass', 'ESRA', 'look', 'superior', 'comparison', 'SRA3', 'however', 'operation', 'justified', 'worst', 'environmental', 'constraints'].*

Taking the target word vector for *"compared"*, we computed the cosine similarity between each word in the comment and the *"compared"* target word vector using the pre-trained GloVe embeddings. This procedure returned a cosine similarity greater than our 0.6 threshold between the word *"comparison"* in the text and the target word vector, "*compared".* Therefore, the dummy variable *Compared_decision* took a value of 1. If instead, there were no words with a cosine similarity higher than our threshold, the dummy variable, *Compared_decision* would take a value of 0.

[Table 10 about here]

4.2.4. Word Embedding Results on Decision Strategy Usage by Evaluator Expertise Type

In Table 11, we report the mean (standard deviation) of each dummy variable corresponding to the probability of detecting each decision strategy in the open-text comments for novelty and feasibility, by *Evaluator expertise type*, Our main objective is to examine how the evaluators applied these strategies to evaluate a solution's quality and whether an evaluator's expertise in the domain shaped their use of the four decision strategies. We use linear probability models (LPMs) to examine the probability of using each decision strategy on *Evaluator expertise type* according to the following regression models for novelty in (4) and feasibility in (5):

$$Quality\ Rating_{ij} = \beta_0 + \beta_1 Evaluator\ Expertise\ Type_i + \beta_2 Novelty\ decision\ strategy_{ij} + \beta_3 Evaluator\ Expertise\ Type_i \cdot Novelty\ decision\ strategy_{ij} + \varepsilon_{ij} \tag{4}$$

$$Quality\ Rating_{ij} = \beta_0 + \beta_1 Evaluator\ Expertise\ Type_i + \beta_2 Feasibility\ decision\ strategy_{ij} + \beta_3 Evaluator\ Expertise\ Type_i \cdot Feasibility\ decision\ strategy_{ij} + \varepsilon_{ij} \tag{5}$$

[Table 11 about here]

In Tables 12 and 13, we present our results examining the evaluators' decision strategies. Table 12 presents the estimated relationships between *Choose the best* (Models 1-4), and *Best compared to others* (Models 5-8). Table 13 presents the estimated relationships between *Avoid the worst* (Models 1-4) and *Avoiding risky solutions* (Models 5-8). For both Tables 12 and 13, Models 1-2 and 5-6 focus on the decision strategies used to come up with the novelty rating for each solution, while Models 3-4 and 7-8 focus on the decision strategies used to come up with the feasibility rating for each solution.

Turning to Table 12, we observe in Model 1 that using a *Choose the best* decision strategy to evaluate a solution's novelty has a positive effect on solution quality (Model 1: 0.551, $p < 0.01$). Turning to the interaction terms between *Evaluator expertise type* x *Novelty_best*, we observe that the coefficients are negative and significant for *HR screen* (Model 2: -0.741, $p < 0.05$) and negative but not significant for *Skills test screen* (Model 2: -0.539, *ns*). This suggests that compared to non-experts, domain experts are less likely to indicate that a solution is high in quality when it is a highly novel design. The patterns for feasibility are similar. We observe in Model 3 that the *Choose the best* design strategy to evaluate a solution's feasibility has a positive effect on the solution's quality rating (Model 3: 0.788, $p < 0.01$). In Model 4 we observe that the interaction terms between *Feasibility_best* and *Evaluator expertise type* is negative for the *HR screen* (Model 4: -0.339, *ns*) and negative and marginally significant for the *Skills test screen* (Model 4: -0.601, $p < 0.10$). Based on the results in Models 1-4, we have some evidence to suggest that non-experts (i.e., the *Unscreened* evaluators) are more likely than experts to apply a *Choose the Best* decision strategy to arrive at their evaluations of solution quality.

Next, examining the *Compared to others* decision strategy, in Models 5-8, first, we find no significant effect of *Novelty_compared* on a solution's quality rating (Model 5: 0.101, *ns*). Similarly, the interaction terms between *Novelty_compared* and *Evaluator expertise type* are both positive but not significant in Model 6 (HR screen: 0.0916, *ns*; Skills test screen: 0.282, *ns*). This suggests that we do not have evidence to support that expert evaluators are more likely than non-experts to arrive at their ratings of solution quality by comparing a solution's novelty to other designs. That said, we find different patterns by evaluator expertise for *Feasibility_compared* and a solution's quality. In Model 7, we find a positive and significant effect of *Feasibility_compared* on solution quality (Model 7: 0.390, $p < 0.05$). Then, in Model 8, although the interaction term between *HR screen* x *Feasibility_compared* is not significant, the interaction terms is positive and significant for *Skills test screen* x *Feasibility_compared* (Model 8: 0.678,

p < 0.05). Thus, we find some evidence to support that the domain experts who passed the *Skills test screen* rate a solution's quality higher when they perceive it to be more feasible *compared* to other solutions.

[Table 12 about here]

Next, we turn to Table 13 for the *Avoid the worst* and *Avoiding risk* strategies. In Model 1, we observe that there is a negative but not significant relationship between avoiding the worst or least novel solution, and a solution's quality (Model 1: -0.471, *ns*). Similarly, in Model 2, the interaction term between *Novelty_worst* and *Evaluator expertise type* is negative but not significant (HR screen: -1.201, *ns*; Skills test screen: -0.562, *ns*). Turning to Model 3, we find that the coefficient for *Feasibility_worst* is negative and marginally significant (Model 3: -0.522, p < 0.10), suggesting that decision strategies that avoid the worst or least feasible solutions are associated with higher quality solutions. That said, the coefficients for the interaction term between *Evaluator expertise type* x *Feasibility_worst* are not significant in Model 4 (HR screen: 0.289, *ns*; Skills test screen: 0.845, *ns*). Taken together, we do not have reliable evidence to suggest that expert evaluators are more likely than non-experts to make assessments about a solution's quality by using decision strategies that avoid the worst solutions in terms of their novelty or feasibility.

Turning to *Avoiding risk* strategies, we observe that the coefficient for *Novelty_risky* is negative and significant (Model 5: -0.644, p < 0.01), suggesting that risky solutions that are perceived to be more novel, tend to receive lower quality ratings. Turning to Model 6, the interaction term between the *HR screen* x *Novelty_risky* is negative but not significant (Model 6: -0.0226, *ns*) and is negative and significant between *Skills test screen* x *Novelty_risky* (Model 6: -1.346, p < 0.05). This suggests that compared to non-experts, the *Skills test screen* evaluators are more likely to rate solutions as higher in quality when they are less novel and less risky. In terms of risk and feasibility, in Model 7 we observe a negative effect of *Feasibility_risky* on solution quality (Model 7: -0.539, p < 0.01). However, neither interaction term with *Evaluator expertise type* is significant in Model 8 (HR screen: -0.0683, *ns*; Skills test screen: 0.128, *ns*).

In summary, we find evidence to support that domain experts and non-experts may use different decision strategies for a solution's novelty and feasibility when coming up with an overall rating of its quality. We find the largest differences in decision strategies between the *Skills test screen* and *Unscreened* evaluator groups. More specifically, we find that while the *Unscreened* evaluators are more likely to use a *Choose the best* decision strategy to arrive at a solution's quality, the *Skills test screen* evaluators are more likely to use a *Compared to others* decision strategy to evaluate a solution's feasibility and an *Avoiding risk* decision strategy to evaluate a solution's novelty. By examining the evaluators' decision strategies, we gain deeper insights into why expert evaluators might exhibit a feasibility preference, in which they overweight solutions that are higher in feasibility but lower in novelty.

[Table 13 about here]

5. Discussion and Conclusion

This paper reports on an experiment designed to investigate how domain expertise is systematically related to the relative importance of novelty and feasibility during evaluations of R&D solutions in technical domains. We conducted this field experiment by layering on an evaluation process on top of an existing innovation challenge series at NASA and collaborating with Freelancer.com to recruit a large number of evaluators with different levels and types of expertise in the domain area to evaluate an array of solution designs. We collected intricate background information on the evaluators, including demographic data and a skills test assessment, which were used for identifying expertise in the domain. This created exogenous variation between the evaluator's domain expertise and the problems they were assigned to evaluate and facilitated causal estimates of domain expertise and evaluation scores, while holding characteristics of the evaluators and solutions constant.

We report several noteworthy patterns. First, our experimental results show that domain experts have a *feasibility preference*, and are more likely to judge solutions as higher in quality when they are more feasible but less novel, suggesting that they perceive novelty and feasibility as tradeoffs in a solution's design. Second, we observe that these patterns are moderated by the complexity of the problem, in which the perceived tradeoff between novelty and feasibility is larger for more complex problems that draw upon multiple domains and are more difficult to solve and hence, evaluate for their "quality". Third, using text analysis to examine the qualitative comments associated with the evaluators' judgments, we observe that domain experts use different decision strategies for assessing the novelty and feasibility of solutions. We find that domain experts pay closer attention to the feasibility of a solution, exhibiting patterns indicative of deeper information processing of a solution's feasibility, but find no such differences by evaluator expertise for a solution's novelty. Leveraging word embedding models, we find that experts identify high quality solutions by comparing across solutions to identify the *most feasible alternative*, while *avoiding risky* but novel solution designs. This suggests that experts and non-experts use different decision heuristics to judge the novelty and feasibility of solutions. Our findings are consistent with the notion that deep domain knowledge is most relevant for judging the feasibility of a given solution, because it is an objective criterion that draws directly on an individual's competence and experience in the domain. This is reinforced by the technical nature of the robotics domain, in which the asymmetric risk-reward payoffs of avoiding a failed product, device or system creates an ambiguity aversion (Fox and Tversky 1995) that negates the importance of creating novel designs that have are associated with uncertainty and higher risk of failure.

Our study offers a number of insights that point to future avenues of work. First, our study provides clarity into the relationships between novelty, feasibility and expert evaluations of quality. Prior work on evaluation processes suggests that experts exhibit an anti-novelty bias during project evaluation and selection processes, in which evaluators across a range of scientific institutions and innovative organizations tend to discount novel ideas (Boudreau et al. 2016, Mueller et al. 2012, Uzzi et al. 2013, Wang et al. 2017)—rewarding ideas with intermediate levels of novelty over those that are highly novel (Boudreau et al. 2016, Criscuolo et al. 2017). Within technical organizations, we suggest that the greater potential reward of pursuing novel ideas may run counter to the need for certainty in outcomes—skewing domain experts towards highly feasible but less novel solutions. In these domains, the need to implement and produce finished goods, services, and products on time makes novelty or originality more of an afterthought during evaluation processes. Consequently, our results indicate that the magnitude of the anti-novelty bias may depend on the risk-reward tradeoff associated with failures compared to upside potential.

These findings are of particular relevance to the evaluation of innovative R&D projects and solutions. Most evaluation processes draw on individuals with deep knowledge and expertise in the domain; the general belief is that familiarity and competence in the domain are essential for making informed judgments about a potential project or solution's quality (Chase and Simon 1973, Chi et al. 1981, Li 2017, Shanteau 1992b). Despite the role of expertise in evaluations, few studies have attempted to make causal inferences of the role of "distance" in knowledge or technological space to the domain area and evaluation outcomes (see Boudreau et al. 2016 and Li 2017 for exceptions). We contribute to this literature by examining how domain expertise affects the way that evaluators sample or "see" information about a solution's novelty and feasibility when evaluating technical solutions. In this respect, we show that expert evaluators oversample on a solution's feasibility and have a strong *feasibility preference*. These findings have implications for how managers recruit experts to evaluate new R&D ideas. Although managers tend to staff evaluation and selection processes with experts in the domain, we show that high expertise is likely to lead to selection criteria that filter out novel ideas even before they can be considered. Given that these types of decisions can affect a firm's strategic direction, it raises the question of whether managers should aim to deliberately increase the diversity of their evaluator pools by broadening the distance between an evaluator's expertise and the focal domain of the idea or solution. In this regard, whereas a large literature on the scientific peer review process points to the downsides of "noise" in evaluation scores and poor interrater reliability among experts (Cole and Simon 1981, Pier et al. 2018, Rothwell and Martyn 2000), we suggest that inserting deliberate variation in domain distance can encourage consideration of other solutions with alternative risk-reward payoff structures.

Second, our findings reveal that the source of one's expertise within the domain is a critical factor influencing how evaluators value novelty and feasibility in their evaluations of quality. Although we found that both types of expert evaluators (HR screen or prior training/background in domain and Skills screen or high proficiency in domain) exhibited a feasibility preference for highly feasible and less novel solutions, there are some nuanced differences between the two groups. It is worthwhile to note that we found that solution feasibility was even more critical to evaluators demonstrating high proficiency in the domain, even though their average expertise was more distant from the domain of robotics design than evaluators with direct work experience and training in robotics.

Aside from our experts, the pool of unscreened evaluators were "less-expert" but by no means novices in the domain. Despite broadcasting the call to anyone who deems themselves qualified to evaluate the solutions, the registrants who took up the call still straddled the boundary of robotics design as being within their field of expertise. Similar to our expert population, our unscreened evaluators were also highly educated and had relevant work experience in technical disciplines—but in science and technology more broadly. Since there is no objective ground truth, ex ante measure of "true" quality when evaluating R&D solutions, this further suggests that future investigation into how different sources of expertise (e.g., disciplinary versus skills or another dimension) that leverage alternative dimensions of common knowledge overlap to the domain, may result in a broader sampling of evaluation criteria when judging solution quality. It remains a topic for future work to investigate how drawing upon evaluators with experience in analogous domains (Franke et al. 2014) or those who straddle the boundary of multiple domains (Dahlander et al. 2016, Gieryn 1983, Jeppesen and Lakhani 2010) might introduce valuable heterogeneity to evaluation and selection outcomes.

Third, our study has implications for the design of evaluation processes for "risky" ideas in technical domains. Our open-text comments suggest that the expert evaluators were more likely to associate more novel ideas with risk compared to less feasible ones. It suggests that standard evaluation formats that base funding and selection decisions on overall merit scores that average or aggregate across different dimensions (Franzoni and Stephan 2021) may lead to biased outcomes that reflect the preferences of the evaluator pool. These systematic biases cannot be removed even through efforts to draw upon a larger number of evaluators of the same expertise type (Budescu and Chen 2015). A plausible path would be to develop separate processes for evaluating solution feasibility and novelty that would then allow an expert authority (overseeing the evaluation process) to make choices among alternative projects depending on the relative degree of risk in their R&D portfolio. This is consistent with prior findings suggesting that evaluators' risk preferences are affected under conditions that require them to make comparisons or tradeoffs between two tasks for which they hold differing amounts of expertise (Fox and Tversky 1995). Following this logic, it then becomes a question of how to best assign evaluators to different components of the evaluation process. Whereas domain experts potentially have an informational advantage over non-experts in assessing a technical solution's feasibility, there is limited evidence indicating that experts are any more qualified to evaluate a solution's novelty. In fact, from a bounded rationality perspective, there is strong evidence to suggest that experts' mental maps breakdown when applied to new areas (Camerer and Johnson 1991, Simon 1957). Rather a promising avenue might be to leverage less-expert evaluators from adjacent domains who could potentially offer a more objective assessment of a solution's novelty. Ultimately, efforts to reconfigure the evaluation process—paired with appropriate incentives that reward some strategic risk-taking behavior may lead to exploration and the pursuit of radical innovations.

Overall, this study provides insight into how expertise shapes decision criteria and evaluation outcomes of technical R&D problems and offers innovative organizations a deeper understanding of the tradeoffs between novelty and feasibility associated with expertise and evaluations.

6. References

Abernathy WJ (1976) The Productivity Dilemma; Roadblock to Innovation in the Automobile Industry.

Abernathy WJ, Clark KB (1985) Innovation: Mapping the winds of creative destruction. *Res. Policy* 14(1):3–22.

Abernathy WJ, Utterback JM (1978) Patterns of industrial innovation. *Technol. Rev.* 80(7):40–47.

Adamson RE (1952) Functional fixedness as related to problem solving: a repetition of three experiments. *J. Exp. Psychol.* 44(4):288.

Allen TJ (1977) Managing the flow of technology: Technology transfer and the dissemination of technological information within the R&D organisation. *Camb. MA Mass. Inst. Technol.*

Amabile TM (1983) Brilliant but cruel: Perceptions of negative evaluators. *J. Exp. Soc. Psychol.* 19(2):146–156.

Amabile TM (1996) Creativity and innovation in organizations.

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

Audia PG, Goncalo JA (2007) Past success and creativity over time: A study of inventors in the hard disk drive industry. *Manag. Sci.* 53(1):1–15.

Azoulay P, Graff Zivin JS, Li D, Sampat BN (2019) Public R&D investments and private-sector patenting: evidence from NIH funding rules. *Rev. Econ. Stud.* 86(1):117–152.

Azoulay P, Li D (2020) *Scientific Grant Funding* (National Bureau of Economic Research).

Baer M (2012) Putting creativity to work: The implementation of creative ideas in organizations. *Acad. Manage. J.* 55(5):1102–1119.

Baum JR, Bird BJ (2010) The successful intelligence of high-growth entrepreneurs: Links to new venture growth. *Organ. Sci.* 21(2):397–412.

Bengio Y, Ducharme R, Vincent P, Jauvin C (2003) A neural probabilistic language model. *Neural Probabilistic Lang. Model* 3:1137–1155.

Benner MJ, Tushman ML (2003) Exploitation, exploration, and process management: The productivity dilemma revisited. *Acad. Manage. Rev.* 28(2):238–256.

Berg JM (2016) Balancing on the creative highwire: Forecasting the success of novel ideas in organizations. *Adm. Sci. Q.* 61(3):433–468.

Berg JM (2022) One-Hit Wonders versus Hit Makers: Sustaining Success in Creative Industries. *Adm. Sci. Q.*:00018392221083650.

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

Blumenstock JE (2008) Size matters: word count as a measure of quality on wikipedia. *Proc. 17th Int. Conf. World Wide Web*. 1095–1096.

Boudreau KJ, Guinan EC, Lakhani KR, Riedl C (2016) Looking across and looking beyond the knowledge frontier: Intellectual distance, novelty, and resource allocation in science. *Manag. Sci.* 62(10):2765–2783.

Boudreau KJ, Lacetera N, Lakhani KR (2011) Incentives and problem uncertainty in innovation contests: An empirical analysis. *Manag. Sci.* 57(5):843–863.

Bower JL (1972) *Managing the resource allocation process: A study of corporate planning and investment* (Irwin Homewood).

Budescu DV, Chen E (2015) Identifying expertise to extract the wisdom of crowds. *Manag. Sci.* 61(2):267–280.

Burns TE, Stalker GM (1961) The management of innovation.

Camerer CF, Johnson EJ (1991) *The process-performance paradox in expert judgment* (Cambridge University Press, Cambridge, MA).

Cannon MD, Edmondson AC (2005) Failing to learn and learning to fail (intelligently): How great organizations put failure to work to innovate and improve. *Long Range Plann.* 38(3):299–319.

Chai S, Doshi AR, Silvestri L (2021) How Catastrophic Innovation Failure Affects Organizational and Industry Legitimacy: The 2014 Virgin Galactic Test Flight Crash. *Organ. Sci.*

Chase WG, Simon HA (1973) Perception in chess. *Cognit. Psychol.* 4(1):55–81.

Chi MT, Glaser R, Rees E (1981) *Expertise in problem solving.* (Pittsburgh Univ PA Learning Research and Development Center).

Christensen CM, Bower JL (1996) Customer power, strategic investment, and the failure of leading firms. *Strateg. Manag. J.* 17(3):197–218.

Cole S, Simon GA (1981) Chance and consensus in peer review. *Science* 214(4523):881–886.

Criscuolo P, Dahlander L, Grohsjean T, Salter A (2017) Evaluating novelty: The role of panels in the selection of R&D projects. *Acad. Manage. J.* 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. *Organ. Sci.*

Csikszentmihalyi M (1999) 16 implications of a systems perspective for the study of creativity. *Handb. Creat.* 313.

Cyert RM, March JG (1963) A behavioral theory of the firm. *Englewood Cliffs NJ* 2(4):169–187.

Dahlander L, O'Mahony S, Gann DM (2016) One foot in, one foot out: how does individuals' external search breadth affect innovation outcomes? *Strateg. Manag. J.* 37(2):280–302.

Dane E (2010) Reconsidering the trade-off between expertise and flexibility: A cognitive entrenchment perspective. *Acad. Manage. Rev.* 35(4):579–603.

Danneels E (2002) The dynamics of product innovation and firm competences. *Strateg. Manag. J.* 23(12):1095–1121.

Dimitriadis S, Koning R (2022) Social skills improve business performance: evidence from a randomized control trial with entrepreneurs in Togo. *Manag. Sci.*

Dosi G (1982) Technological paradigms and technological trajectories: a suggested interpretation of the determinants and directions of technical change. *Res. Policy* 11(3):147–162.

Douglas M, Wildavsky A (1983) Risk and culture. *Risk Cult.* (University of California press).

Duncker K (1945) On problem-solving. *Psychol. Monogr.* 58(5):i.

Einhorn HJ (1974) Expert judgment: Some necessary conditions and an example. *J. Appl. Psychol.* 59(5):562.

Eisenhardt KM, Martin JA (2000) Dynamic capabilities: what are they? *Strateg. Manag. J.* 21(10-11):1105–1121.

Eisenhardt KM, Tabrizi BN (1995) Accelerating adaptive processes: Product innovation in the global computer industry. *Adm. Sci. Q.*:84–110.

Ferguson JP, Carnabuci G (2017) Risky recombinations: Institutional gatekeeping in the innovation process. *Organ. Sci.* 28(1):133–151.

Fleming L (2001) Recombinant uncertainty in technological search. *Manag. Sci.* 47(1):117–132.

Florman SC (2014) *The civilized engineer* (St. Martin's Griffin).

Ford CM (1996) A theory of individual creative action in multiple social domains. *Acad. Manage. Rev.* 21(4):1112–1142.

Fox CR, Tversky A (1995) Ambiguity aversion and comparative ignorance. *Q. J. Econ.* 110(3):585–603.

Franke N, Poetz MK, Schreier M (2014) Integrating problem solvers from analogous markets in new product ideation. *Manag. Sci.* 60(4):1063–1081.

Franzoni C, Stephan P (2021) *Uncertainty and Risk-Taking in Science: Meaning, Measurement and Management* (National Bureau of Economic Research).

Franzoni C, Stephan P, Veugelers R (2021) *Funding Risky Research* (National Bureau of Economic Research).

Gallo S, Thompson L, Schmaling K, Glisson S (2018) Risk evaluation in peer review of grant applications. *Environ. Syst. Decis.* 38(2):216–229.

Gieryn TF (1983) Boundary-work and the demarcation of science from non-science: Strains and interests in professional ideologies of scientists. *Am. Sociol. Rev.*:781–795.

Gigerenzer G, Goldstein DG (1996) Reasoning the fast and frugal way: models of bounded rationality. *Psychol. Rev.* 103(4):650.

Gigone D, Hastie R (1997) Proper analysis of the accuracy of group judgments. *Psychol. Bull.* 121(1):149.

Gilbert RJ, Newbery DM (1982) Preemptive patenting and the persistence of monopoly. *Am. Econ. Rev.*:514–526.

Goldstein WM, Weber EU (1995) Content and discontent: Indications and implications of domain specificity in preferential decision making. *Psychol. Learn. Motiv.* (Elsevier), 83–136.

Hardiman PT, Dufresne R, Mestre JP (1989) The relation between problem categorization and problem solving among experts and novices. *Mem. Cognit.* 17(5):627–638.

Helfat CE, Peteraf MA (2015) Managerial cognitive capabilities and the microfoundations of dynamic capabilities. *Strateg. Manag. J.* 36(6):831–850.

Henderson R (1993) Underinvestment and incompetence as responses to radical innovation: Evidence from the photolithographic alignment equipment industry. *RAND J. Econ.*:248–270.

Henderson R, Clark KB (1990) Architectural innovation: The reconfiguration of existing product technologies and the failure of established firms. *Adm. Sci. Q.*:9–30.

Hennig A, Topcu TG, Szajnfarber Z (2022) So You Think Your System Is Complex?: Why and How Existing Complexity Measures Rarely Agree. *J. Mech. Des.* 144(4).

Hinds PJ (1999) The curse of expertise: The effects of expertise and debiasing methods on prediction of novice performance. *J. Exp. Psychol. Appl.* 5(2):205.

Hinds PJ, Patterson M, Pfeffer J (2001) Bothered by abstraction: The effect of expertise on knowledge transfer and subsequent novice performance. *J. Appl. Psychol.* 86(6):1232.

Holmfeld JD (1970) *Communication behavior of scientists and engineers* (Case Western Reserve University).

Jeppesen LB, Lakhani KR (2010) Marginality and problem-solving effectiveness in broadcast search. *Organ. Sci.* 21(5):1016–1033.

Johnson EJ (1988) Expertise and decision under uncertainty: Performance and process. *Nat. Expert.*:209–228.

Jordan KN, Sterling J, Pennebaker JW, Boyd RL (2019) Examining long-term trends in politics and culture through language of political leaders and cultural institutions. *Proc. Natl. Acad. Sci.* 116(9):3476–3481.

Kahneman D, Klein G (2009) Conditions for intuitive expertise: a failure to disagree. *Am. Psychol.* 64(6):515.

Kahneman D, Slovic SP, Slovic P, Tversky A (1982) *Judgment under uncertainty: Heuristics and biases* (Cambridge university press).

Kaplan S, Vakili K (2015) The double-edged sword of recombination in breakthrough innovation. *Strateg. Manag. J.* 36(10):1435–1457.

Kavadias S, Sommer SC (2009) The effects of problem structure and team diversity on brainstorming effectiveness. *Manag. Sci.* 55(12):1899–1913.

Kornish LJ, Ulrich KT (2011) Opportunity spaces in innovation: Empirical analysis of large samples of ideas. *Manag. Sci.* 57(1):107–128.

Kornish LJ, Ulrich KT (2014) The importance of the raw idea in innovation: Testing the sow's ear hypothesis. *J. Mark. Res.* 51(1):14–26.

Krieger JL, Li D, Papanikolaou D (2021) Missing novelty in drug development. *Rev. Financ. Stud.* (Forthcoming).

Lag S (2022) Reusable rockets: revolutionizing access to outer space. *Technol. Outlook 2030*.

Lane J, Teplitskiy M, Gray G, Ranu H, Menietti M, Guinan E, Lakhani KR (2021) Conservatism Gets Funded? A Field Experiment on the Role of Negative Information in Novel Project Evaluation. *Manag. Sci. Forthcom.*

Layton Jr ET (1974) Technology as knowledge. *Technol. Cult.*:31–41.

Leonard-Barton D (1992) Core capabilities and core rigidities: A paradox in managing new product development. *Strateg. Manag. J.* 13(S1):111–125.

Levinthal DA, March JG (1993) The myopia of learning. *Strateg. Manag. J.* 14(S2):95–112.

Li D (2017) Expertise versus Bias in Evaluation: Evidence from the NIH. *Am. Econ. J. Appl. Econ.* 9(2):60–92.

Luchins AS (1942) Mechanization in problem solving: The effect of einstellung. *Psychol. Monogr.* 54(6):i.

MacCrimmon KR, Wehrung D, Stanbury WT (1988) *Taking risks* (Simon and Schuster).

March JG (1991) Exploration and exploitation in organizational learning. *Organ. Sci.* 2(1):71–87.

Maritan CA, Lee GK (2017) *Resource allocation and strategy* (SAGE Publications Sage CA: Los Angeles, CA).

Meehl PE (1954) Clinical versus statistical prediction: A theoretical analysis and a review of the evidence.

Mollick E, Nanda R (2016) Wisdom or madness? Comparing crowds with expert evaluation in funding the arts. *Manag. Sci.* 62(6):1533–1553.

Moran P, Ghoshal S (1999) Markets, firms, and the process of economic development. *Acad. Manage. Rev.* 24(3):390–412.

Moreau CP, Lehmann DR, Markman AB (2001) Entrenched knowledge structures and consumer response to new products. *J. Mark. Res.* 38(1):14–29.

Mueller JS, Melwani S, Goncalo JA (2012) The bias against creativity: Why people desire but reject creative ideas. *Psychol. Sci.* 23(1):13–17.

Nelson RR, Winter SG (1982) The Schumpeterian tradeoff revisited. *Am. Econ. Rev.* 72(1):114–132.

Noda T, Bower JL (1996) Strategy making as iterated processes of resource allocation. *Strateg. Manag. J.* 17(S1):159–192.

Pennebaker JW, Boyd RL, Jordan K, Blackburn K (2015) *The development and psychometric properties of LIWC2015*

Pennebaker JW, Mehl MR, Niederhoffer KG (2003) Psychological aspects of natural language use: Our words, our selves. *Annu. Rev. Psychol.* 54(1):547–577.

Pennington J, Socher R, Manning CD (2014) Glove: Global vectors for word representation. *Proc. 2014 Conf. Empir. Methods Nat. Lang. Process. EMNLP*. 1532–1543.

Pier EL, Brauer M, Filut A, Kaatz A, Raclaw J, Nathan MJ, Ford CE, Carnes M (2018) Low agreement among reviewers evaluating the same NIH grant applications. *Proc. Natl. Acad. Sci.* 115(12):2952–2957.

Reitzig M, Sorenson O (2013) Biases in the selection stage of bottom-up strategy formulation. *Strateg. Manag. J.* 34(7):782–799.

Rettinger DA, Hastie R (2001) Content effects on decision making. *Organ. Behav. Hum. Decis. Process.* 85(2):336–359.

Rosenberg N (1972) Factors affecting the diffusion of technology. *Explor. Econ. Hist.* 10(1):3.

Rosenkopf L, Nerkar A (2001) Beyond local search: boundary-spanning, exploration, and impact in the optical disk industry. *Strateg. Manag. J.* 22(4):287–306.

Ross J (2018) A Brief Recap of Reusable Rockets. *ZLSA Des.*

Rothwell PM, Martyn CN (2000) Reproducibility of peer review in clinical neuroscience: Is agreement between reviewers any greater than would be expected by chance alone? *Brain* 123(9):1964–1969.

Schumpeter JA (1942) *Capitalism, socialism and democracy* (Routledge).

Schwartz B, Ward A, Monterosso J, Lyubomirsky S, White K, Lehman DR (2002) Maximizing versus satisficing: happiness is a matter of choice. *J. Pers. Soc. Psychol.* 83(5):1178.

Shanteau J (1992a) Competence in experts: The role of task characteristics. *Organ. Behav. Hum. Decis. Process.* 53(2):252–266.

Shanteau J (1992b) How much information does an expert use? Is it relevant? *Acta Psychol. (Amst.)* 81(1):75–86.

Simon HA (1955) A behavioral model of rational choice. *Q. J. Econ.* 69(1):99–118.

Simon HA (1957) Models of man; social and rational.

Simon HA (1978) Information-processing theory of human problem solving. *Handb. Learn. Cogn. Process.* 5:271–295.

Simon HA (1987) Making management decisions: The role of intuition and emotion. *Acad. Manag. Perspect.* 1(1):57–64.

Simon HA (2019) *The Sciences of the Artificial, reissue of the third edition with a new introduction by John Laird* (MIT press).

Simonton DK (1999) Creativity as blind variation and selective retention: Is the creative process Darwinian? *Psychol. Inq.*:309–328.

Sitkin SB, Pablo AL (1992) Reconceptualizing the determinants of risk behavior. *Acad. Manage. Rev.* 17(1):9–38.

Sitkin SB, Weingart LR (1995) Determinants of risky decision-making behavior: A test of the mediating role of risk perceptions and propensity. *Acad. Manage. J.* 38(6):1573–1592.

Sommer SC, Bendoly E, Kavadias S (2020) How do you search for the best alternative? Experimental evidence on search strategies to solve complex problems. *Manag. Sci.* 66(3):1395–1420.

Staw BM, Ross J (1987) Knowing when to pull the plug. *Harv. Bus. Rev.* 65(2):68–74.

Szajnfarber Z, Zhang L, Mukherjee S, Crusan J, Hennig A, Vrolijk A (2020) Who is in the Crowd? Characterizing the capabilities of prize competition competitors. *IEEE Trans. Eng. Manag.*

Tausczik YR, Pennebaker JW (2010) The psychological meaning of words: LIWC and computerized text analysis methods. *J. Lang. Soc. Psychol.* 29(1):24–54.

Taylor A, Greve HR (2006) Superman or the fantastic four? Knowledge combination and experience in innovative teams. *Acad. Manage. J.* 49(4):723–740.

Teece DJ, Pisano G, Shuen A (1997) Dynamic capabilities and strategic management. *Strateg. Manag. J.* 18(7):509–533.

Tetlock PE (2009) *Expert political judgment* (Princeton University Press).

Thomke SH (1998) Managing experimentation in the design of new products. *Manag. Sci.* 44(6):743–762.

Tripsas M (1997) Unraveling the process of creative destruction: Complementary assets and incumbent survival in the typesetter industry. *Strateg. Manag. J.* 18(S1):119–142.

Tushman ML, Anderson P (1986) Technological discontinuities and organizational environments. *Adm. Sci. Q.*:439–465.

Tversky A, Kahneman D (1974) Judgment under uncertainty: Heuristics and biases. *science* 185(4157):1124–1131.

Uzzi B, Mukherjee S, Stringer M, Jones B (2013) Atypical combinations and scientific impact. *Science* 342(6157):468–472.

Vincenti WG (1990) *What engineers know and how they know it* (Baltimore: Johns Hopkins University Press).

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

Weber EU, Johnson EJ (2009) Mindful judgment and decision making. *Annu. Rev. Psychol.* 60:53–85.

Weitzman ML (1998) Recombinant growth. *Q. J. Econ.* 113(2):331–360.

Wicht A, Szajnfarber Z (2014) Portfolios of promise: A review of R&D investment techniques and how they apply to technology development in space agencies. *Space Policy* 30(2):62–74.

Woolnough BE (1991) *The Making of Engineers and Scientists: An Enquiry Into the Factors Affecting Schools Success in Producing Engineers and Scientists* (Oxford University Department of Educational Studies).

Table 1. Number of Evaluator-Solution Pairs by Challenge and Evaluator Expertise Type (N = 3,869)

Challenge	# Of Solutions	Boundary-spanner	Expert	Expert	Mean (s.d.) per
		(Unscreened)	(HR screen)	(Skills screen)	solution
HMSA	6	125	126	185	24.222 (5.047)
MDC	14	145	115	135	9.405 (2.142)
MIS	11	164	125	162	13.667 (2.869)
PSA	6	125	115	190	23.889 (5.860)
RASA	6	140	130	141	22.833 (1.543)
SAM	12	170	140	185	13.750 (2.644)
SDM	18	141	100	142	7.093 (2.174)
SPAM	12	139	145	170	12.611 (2.664)
SRA	16	139	125	150	12.611 (2.665)
Total	101	1,288	1,121	1,460	12.769 (6.319)

Table 2. Summary Statistics on Alternative Measures of Evaluator Expertise in Domain

	Overall	Non-Expert	Expert	Expert	Chi-Sq/ANOVA Test
		(Unscreened)	(HR Screen)	(Skills Screen)	
	(N = 374)	(N = 125)	(N = 109)	(N = 140)	
HR screen (work experience)	1.659 (3.067)	0.380 (2.050)	4.056 (3.544)	0.946 (2.297)	F(2,369) = 63.72***
Skills test screen (skills test score)	10.366 (3.889)	7.128 (3.245)	9.620 (3.188)	13.856 (1.060)	F(2.369) = 221.80***
Robotics expertise distance	2.589 (1.362)	3.528 (1.411)	1.917 (1.033)	2.267 (1.053)	χ^2(10,374) = 120.52***
Distance to roboticist discipline	1.788 (0.966)	2.496 (0.956)	1.000 (0.000)	1.763 (0.865)	χ^2(10,374) = 177.14***

Note: HR screen and Skills test screen were the two criteria used to identify evaluator expertise in the domain. To meet the HR screen threshold, evaluators needed two or more years of work experience in the robotics domain. To meet the skills test screen threshold, evaluators needed to score a 13 or more out of 17 on the skills test.

Table 3. Summary Statistics on Evaluator Demographics (N = 374)

	Overall	Non-Expert	Expert	Expert	Chi-Sq Test
		(Unscreened)	(HR Screen)	(Skills Screen)	
	(N = 374)	(N = 125)	(N = 109)	(N = 140)	
Female	0.154 (0.361)	0.231 (0.423)	0.126 (0.333)	0.105 (0.308)	χ^2(2,374) = 14.165, p = 0.007
Age range	2.029 (1.105)	1.735 (0.986)	2.287 (1.267)	2.104 (1.028)	χ^2(14,374) = 29.954, p = 0.008
Bachelors	0.761 (0.427)	0.696 (0.462)	0.787 (0.411)	0.799 (0.403)	χ^2(2,374) = 4.381, p = 0.112
Masters	0.250 (0.434)	0.152 (0.360)	0.287 (0.454)	0.309 (0.464)	χ^2(10,374) = 9.804, p = 0.007
USA	0.067 (0.251)	0.064 (0.246)	0.111 (0.316)	0.036 (0.187)	χ^2(4,374) = 6.545, p = 0.162

Note: Age range levels are 0 = Under 18 or prefer not to say, 1 = 18 - 24, 2 = 25 - 34, 3 = 35 - 44, 4 = 45 - 54, 5 = 55 - 64, 6 = 65 or olde

Table 4. Correlation Table of Main Variables (N = 3,869)

	Variable |	Mean	Std. Dev.	Min	Max	1	2	3	4	5	6	7	8	9	10	11	12
1	Quality rating	4.648	1.811	1	7	1.000											
2	Novelty rating	4.401	1.805	1	7	0.733	1.000										
3	Feasibility rating	4.723	1.790	1	7	0.807	0.649	1.000									
4	Feasibility preference	0.323	1.507	-6	6	0.080	-0.428	0.411	1.000								
5	Eval. expertise type	2.040	0.843	1	3	-0.134	-0.148	-0.110	0.047	1.000							
6	High complexity	0.355	0.478	0	1	0.023	0.053	0.007	-0.055	-0.001	1.000						
7	Novelty confidence	5.613	1.402	1	7	0.130	0.119	0.119	-0.001	0.116	0.020	1.000					
8	Feasibility confidence	5.692	1.362	1	7	0.126	0.111	0.127	0.017	0.125	0.019	0.817	1.000				
9	Female	0.140	0.347	0	1	0.013	0.049	0.003	-0.055	-0.124	0.024	-0.108	-0.109	1.000			
10	Age	1.920	1.213	0	5	-0.136	-0.145	-0.112	0.041	0.169	-0.029	0.120	0.105	-0.076	1.000		
11	USA	0.064	0.246	0	1	0.013	-0.010	0.000	0.012	-0.049	-0.039	0.049	0.047	-0.106	-0.416	1.000	
12	Bachelors	0.764	0.425	0	1	-0.107	-0.078	-0.087	-0.010	0.106	-0.026	0.022	0.037	0.030	0.171	-0.052	1.000
13	Masters	0.245	0.430	0	1	-0.082	-0.094	-0.051	0.053	0.153	-0.060	0.070	0.065	-0.037	0.252	0.048	0.230

ρ > |0.037| significant at p < 0.05 level of significance.

Table 5. Estimated Relationships between Quality Rating, Evaluator Expertise Type, Novelty and Feasibility Ratings

					DV: Solution Quality				
VARIABLES	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8	Model 9
HR screen	-0.356***	-0.123***	-0.254*	-0.618***	-0.535***	-0.628***	-0.614***	-0.616***	-0.524***
	(0.0718)	(0.0400)	(0.130)	(0.144)	(0.140)	(0.144)	(0.144)	(0.158)	(0.143)
Skills test screen	-0.576***	-0.0771**	-0.351***	-0.706***	-0.659***	-0.765***	-0.750***	-0.746***	-0.664***
	(0.0677)	(0.0380)	(0.119)	(0.131)	(0.125)	(0.132)	(0.131)	(0.201)	(0.129)
Novelty rating		0.362***	0.331***	0.362***	0.390***	0.383***	0.379***	0.377***	0.366***
		(0.0167)	(0.0233)	(0.0165)	(0.0315)	(0.0314)	(0.0315)	(0.0385)	(0.0301)
HR screen x Novelty			0.0275		-0.0611	-0.0561	-0.0494	-0.0472	-0.0452
			(0.0247)		(0.0431)	(0.0429)	(0.0431)	(0.0416)	(0.0404)
Skills screen x Novelty			0.0621***		-0.0313	-0.0241	-0.0245	-0.0240	-0.0200
			(0.0231)		(0.0401)	(0.0400)	(0.0402)	(0.0528)	(0.0371)
Feasibility rating		0.578***	0.578***	0.497***	0.479***	0.469***	0.469***	0.468***	0.465***
		(0.0170)	(0.0170)	(0.0257)	(0.0338)	(0.0339)	(0.0339)	(0.0459)	(0.0333)
HR screen x Feasibility				0.101***	0.141***	0.149***	0.141***	0.139**	0.118***
				(0.0263)	(0.0447)	(0.0445)	(0.0444)	(0.0510)	(0.0411)
Skills screen x Feasibility				0.131***	0.151***	0.162***	0.163***	0.162**	0.139***
				(0.0242)	(0.0411)	(0.0411)	(0.0412)	(0.0604)	(0.0388)
Novelty confidence						0.0441	0.0453	0.0464	0.0522
						(0.0297)	(0.0301)	(0.0286)	(0.0315)
Feasibility confidence						0.00306	0.00473	0.00627	0.00277
						(0.0301)	(0.0305)	(0.0208)	(0.0305)
							(0.0375)	(0.0436)	(0.0406)
Constant	4.967***	0.387***	0.538***	0.792***	0.748***	0.581***	0.614***	0.598***	0.672***
	(0.0471)	(0.0597)	(0.104)	(0.117)	(0.114)	(0.121)	(0.162)	(0.147)	(0.131)
Evaluator controls	N	N	N	N	N	N	Y	Y	Y
Challenge FE	N	N	N	N	N	N	N	Y	N
Solution FE	N	N	N	N	N	N	N	N	Y
Observations	3,869	3,869	3,869	3,869	3,869	3,869	3,830	3,830	3,830
R-squared	0.018	0.730	0.730	0.733	0.733	0.734	0.734	0.731	0.696

Robust standard errors in parentheses; *** p<0.01, ** p<0.05, * p<0.1

Table 6. Estimated Relationships between Quality Rating, Evaluator Expertise Type and Feasibility Preference

	DV: Quality Rating					
VARIABLES	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
HR screen	-0.357***	-0.392***	-0.531***	-0.385***	-0.371**	-0.357***
	(0.0717)	(0.0738)	(0.0727)	(0.0742)	(0.115)	(0.0639)
Skills test screen	-0.593***	-0.640***	-0.726***	-0.572***	-0.543***	-0.516***
	(0.0676)	(0.0694)	(0.0684)	(0.0705)	(0.103)	(0.0759)
Feasibility preference	0.104***	0.0100	0.00524	0.0234	0.0283	0.0393
	(0.0187)	(0.0314)	(0.0308)	(0.0315)	(0.0438)	(0.0347)
HR screen x Feas. pref.		0.133***	0.136***	0.115**	0.0995**	0.0500
		(0.0485)	(0.0479)	(0.0491)	(0.0410)	(0.0492)
Skills screen x Feas. pref.		0.148***	0.155***	0.150***	0.150*	0.0980**
		(0.0429)	(0.0424)	(0.0431)	(0.0700)	(0.0437)
Novelty confidence			0.122***	0.140***	0.139**	0.155***
			(0.0373)	(0.0381)	(0.0464)	(0.0395)
Feasibility confidence			0.0949**	0.104***	0.111**	0.0836**
			(0.0380)	(0.0385)	(0.0397)	(0.0378)
Constant	4.941***	4.965***	3.811***	4.363***	4.283***	4.336***
	(0.0476)	(0.0482)	(0.137)	(0.200)	(0.466)	(0.244)
Evaluator controls	N	N	N	Y	Y	Y
Challenge FE	N	N	N	N	Y	N
Solution FE	N	N	N	N	N	Y
Observations	3,869	3,869	3,869	3,830	3,830	3,830
R-squared	0.026	0.029	0.053	0.086	0.088	0.091

Robust standard errors in parentheses; *** p<0.01, ** p<0.05, * p<0.1

Table 7. Estimated Relationships between Quality Rating, Evaluator Expertise Type and Feasibility Preference for Low Complexity Problems

	DV: Quality Rating (Low Complexity Problems)						
VARIABLES	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7
HR screen	-0.390***	-0.385***	-0.413***	-0.590***	-0.470***	-0.446**	-0.443***
	(0.0880)	(0.0877)	(0.0913)	(0.0884)	(0.0913)	(0.131)	(0.0766)
Skills test screen	-0.614***	-0.625***	-0.643***	-0.764***	-0.661***	-0.620***	-0.593***
	(0.0834)	(0.0832)	(0.0868)	(0.0842)	(0.0870)	(0.133)	(0.101)
Feasibility preference		0.0900***	0.0483	0.0312	0.0371	0.0435	0.0673
		(0.0229)	(0.0384)	(0.0373)	(0.0377)	(0.0606)	(0.0437)
HR screen x Feas. pref.			0.0866	0.102*	0.0864	0.0666	0.0204
			(0.0587)	(0.0576)	(0.0582)	(0.0467)	(0.0551)
Skills screen x Feas. pref.			0.0488	0.0702	0.0723	0.0679	0.0229
			(0.0532)	(0.0521)	(0.0526)	(0.0718)	(0.0531)
Novelty confidence				0.129***	0.147***	0.151**	0.186***
				(0.0464)	(0.0470)	(0.0570)	(0.0478)
Feasibility confidence				0.153***	0.155***	0.158**	0.124***
				(0.0480)	(0.0487)	(0.0444)	(0.0437)
Constant	4.960***	4.928***	4.943***	3.454***	4.034***	3.927***	3.904***
	(0.0576)	(0.0584)	(0.0599)	(0.160)	(0.238)	(0.642)	(0.312)
Evaluator controls	N	N	N	N	Y	Y	Y
Challenge FE	N	N	N	N	N	Y	N
Solution FE	N	N	N	N	N	N	Y
Observations	2,506	2,506	2,506	2,506	2,472	2,472	2,472
R-squared	0.021	0.027	0.028	0.072	0.096	0.098	0.110

Robust standard errors in parentheses; *** p<0.01, ** p<0.05, * p<0.1

Table 8. Estimated Relationships between Quality Rating, Evaluator Expertise Type and Feasibility Preference for High Complexity Problems

	DV: Quality Rating (High Complexity Problems)						
VARIABLES	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7
HR screen	-0.299**	-0.318**	-0.340***	-0.389***	-0.188	-0.185	-0.141
	(0.124)	(0.125)	(0.126)	(0.127)	(0.129)	(0.259)	(0.107)
Skills test screen	-0.505***	-0.536***	-0.601***	-0.625***	-0.381***	-0.373*	-0.351***
	(0.116)	(0.116)	(0.116)	(0.116)	(0.122)	(0.103)	(0.101)
Feasibility preference		0.136***	-0.0641	-0.0577	-0.0115	-0.0112	-0.0320
		(0.0327)	(0.0524)	(0.0529)	(0.0578)	(0.0684)	(0.0588)
HR screen x Feas. pref.			0.225***	0.217**	0.177*	0.170	0.119
			(0.0842)	(0.0846)	(0.0910)	(0.0910)	(0.0950)
Skills screen x Feas. pref.			0.361***	0.355***	0.333***	0.339*	0.276***
			(0.0679)	(0.0685)	(0.0728)	(0.0901)	(0.0623)
Novelty confidence				0.0676	0.0786	0.0706	0.0443
				(0.0632)	(0.0653)	(0.0552)	(0.0678)
Feasibility confidence				0.00421	0.0279	0.0450	0.0264
				(0.0604)	(0.0618)	(0.0443)	(0.0671)
Constant	4.982***	4.971***	4.987***	4.605***	5.083***	5.028**	5.270***
	(0.0816)	(0.0828)	(0.0818)	(0.240)	(0.334)	(0.541)	(0.355)
Evaluator controls	N	N	N	N	Y	Y	Y
Challenge FE	N	N	N	N	N	Y	N
Solution FE	N	N	N	N	N	N	Y
Observations	1,363	1,363	1,363	1,363	1,358	1,358	1,358
R-squared	0.013	0.025	0.040	0.042	0.098	0.101	0.094

Robust standard errors in parentheses; *** p<0.01, ** p<0.05, * p<0.1

Table 9. Text Analysis Results on Attentional Focus to Novelty and Feasibility Comments by Evaluator Expertise Type

	Novelty Comments				Feasibility Comments			
	Non-expert (Unscreened)	Expert (HR screen)	Expert (Skills screen)	Pairwise diff. (Tukey)	Non-expert (Unscreened)	Expert (HR screen)	Expert (Skills screen)	Pairwise diff. (Tukey)
Word count	20.7 (22.7)	18.6 (12.6)	19.5 (13.7)	H-U: -2.09*** S-U: -1.24 S-H: 0.85	27.5 (39.6)	26.2 (27.9)	34.3 (52.0)	H-U: -1.25 S-U: 6.78*** S-H: 8.04***
Six letter words	24.3 (14.5)	25.8 (14.5)	26.3 (12.4)	H-U: 1.50** S-U: 1.99*** S-H: 0.49	26.1 (14.6)	28.2 (14.0)	28.5 (11.9)	H-U: 2.09*** S-U: 2.37*** S-H: 0.28
Words per sentence	14.1 (8.21)	13.5 (7.30)	13.8 (8.26)	H-U: -0.56 S-U: -0.21 S-H: 0.35	14.8 (9.31)	13.6 (7.75)	15.2 (9.88)	H-U: -1.21*** S-U: 0.36 S-H: 1.58***
Analytic words	66.6 (33.8)	66.7 (33.7)	68.9 (32.1)	H-U: 0.05 S-U: 2.26 S-H: 2.21	67.7 (32.5)	70.9 (31.2)	71.3 (29.4)	H-U: 3.17** S-U: 3.58*** S-H: 0.41

Note: For the pairwise differences, U = Unscreened, H = HR Screen, and S = Skills test Screen. *** p<0.01, ** p<0.05, * p<0.1

Table 10. Summary of Target Word Vectors Used in Word Embedding Models to Extract Decision Strategies

Decision Rule	Target Word	Synonyms from Comments	Sample Comments
Choose the best	Best	Best, ever, excellent, first, performance, success, winning	Functional analysis is done well and it is best with great effort. It has an excellent description and meets the requirements, you can see the effort made
Best compare to others	Compared	Comparable, compared, comparison, contrast, less, previous, year	This NASA based SAM is highly effective as the Power profile of this design is significantly improved. Moreover, the smart attachment mechanisms of Pack, release and pull is highly effective contrary to earlier designs. In addition, the mass of the system is low enough in contrast with SAM 6,7 and 11.Moreover, the existence of PLC and other MCU's significantly enhance the control characteristics of SAM1. This design idea uses simple gear and links that are normally in use for 50 years, which takes a rating of 3. The use of contact sensors for automation takes the rating of 5, which gives an average rating of 4.
Avoid the worst	Worst	Bad, serious, worse, worst, biggest, terrible, deadly, hardest, deadly, wake	I find the FOA that they pose to find solutions to some problems doing repetitive things bad and they should look for more practical ways of doing them The design is bad, it has no order, the graphics are on several pages, the content is not understood
Avoid risky solutions	Risky	Complicated, difficult, expensive, potentially, problematic, costly, questionable, risky	The design is same old idea with "the bird in the clock" nothing new nothing special, but the use of two complicated systems like this make the product vulnerable The design could be a lot better but with the same SAM system implementation. The worry is how strong the hinges are

Table 11. Summary Statistics on Decision Strategies by Evaluator Expertise Type

Decision Strategy	Non-expert (Unscreened) (N = 1,167)	Expert (HR screen) (N = 987)	Expert (Skills screen) (N = 1,3811)	F-test statistic
Novelty_best	0.040 (0.197)	0.037 (0.190)	0.017 (0.128)	F(2,3535) = 14.452, *p* = 0.001
Novelty_compared	0.031 (0.173)	0.019 (0.137)	0.018 (0.133)	F(2,3535) = 5.353, *p* = 0.069
Novelty_worst	0.007 (0.083)	0.010 (0.100)	0.009 (0.097)	F(2,3535) = 0.769, *p* = 0.681
Novelty_risky	0.013 (0.113)	0.017 (0.130)	0.015 (0.120)	F(2,3535) = 0.713, *p* = 0.700
Feasibility_best	0.040 (0.197)	0.032 (0.177)	0.026 (0.159)	F(2,3535) = 4.056, *p* = 0.132
Feasibility_compared	0.030 (0.171)	0.023 (0.151)	0.023 (0.151)	F(2,3535) = 1.442, *p* = 0.486
Feasibility_worst	0.009 (0.092)	0.009 (0.095)	0.016 (0.125)	F(2,3535) = 3.725, *p* = 0.155
Feasibility _risky	0.033 (0.178)	0.042 (0.200)	0.041 (0.197)	F(2,3535) = 1.517, *p* = 0.468

Table 12. Linear Probability Models of Decision Strategies for Selecting the “Best” Solution by Evaluator Expertise Type

	Choose the Best				Best Compared to Others			
VARIABLES	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8
HR screen	-0.431***	-0.402***	-0.429***	-0.415***	-0.434***	-0.437***	-0.433***	-0.419***
	(0.0789)	(0.0807)	(0.0786)	(0.0806)	(0.0790)	(0.0799)	(0.0791)	(0.0803)
Skills test screen	-0.429***	-0.410***	-0.433***	-0.412***	-0.442***	-0.449***	-0.439***	-0.454***
	(0.0724)	(0.0737)	(0.0722)	(0.0738)	(0.0726)	(0.0732)	(0.0724)	(0.0737)
N_best	0.551***	0.920***						
	(0.162)	(0.212)						
HR screen x N_best		-0.741**						
		(0.377)						
Skills test screen x N_best		-0.539						
		(0.397)						
F_best			0.788***	1.084***				
			(0.148)	(0.203)				
HR screen x F_best				-0.339				
				(0.355)				
Skills test screen x F_best				-0.601*				
				(0.344)				
N_compared					0.101	-0.0133		
					(0.227)	(0.313)		
HR screen x N_compared						0.0916		
						(0.538)		
Skills test screen x N_compared						0.282		
						(0.548)		
F_compared							0.390**	0.322
							(0.170)	(0.248)
HR screen x F_compared								-0.591
								(0.452)
Skills test screen x F_compared								0.678**
								(0.335)
Constant	4.864***	4.848***	4.855***	4.843***	4.885***	4.888***	4.876***	4.878***
	(0.0517)	(0.0524)	(0.0515)	(0.0523)	(0.0517)	(0.0520)	(0.0517)	(0.0523)
Observations	3,092	3,092	3,092	3,092	3,092	3,092	3,092	3,092
R-squared	0.017	0.018	0.021	0.021	0.014	0.014	0.015	0.017

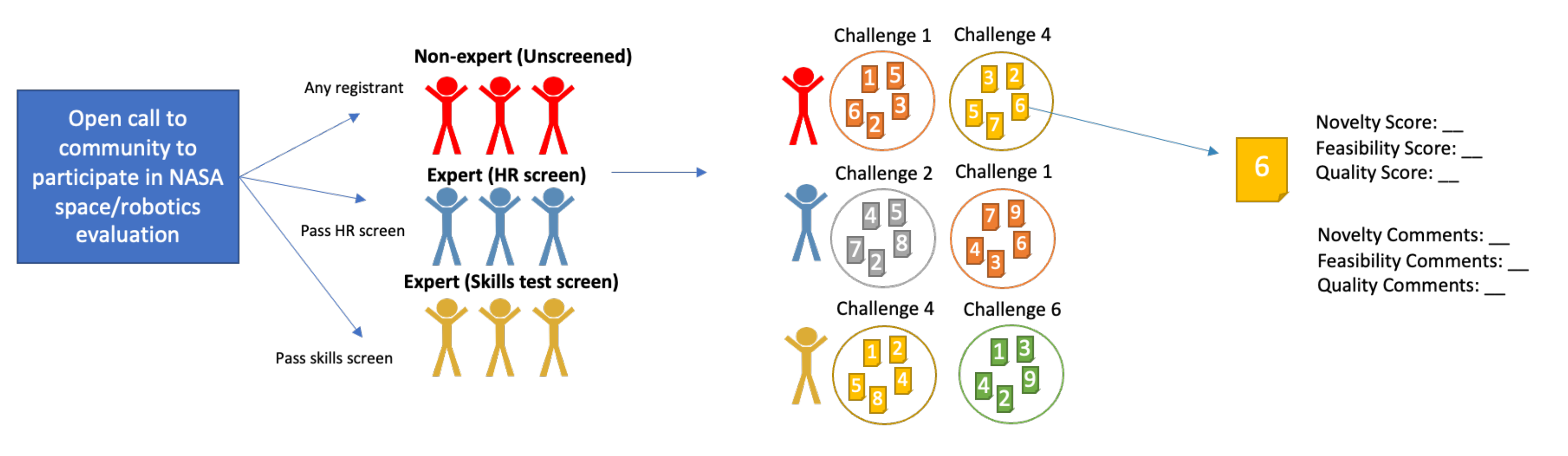
Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table 13. Linear Probability Models of Decision Strategies for Avoiding the "Worst" Solution by Evaluator Expertise Type

	Avoid the Worst				Avoid Risky Solutions			
VARIABLES	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8
HR screen	-0.435***	-0.424***	-0.436***	-0.439***	-0.432***	-0.434***	-0.430***	-0.427***
	(0.0790)	(0.0793)	(0.0789)	(0.0792)	(0.0791)	(0.0799)	(0.0788)	(0.0807)
Skills test screen	-0.442***	-0.438***	-0.441***	-0.451***	-0.442***	-0.423***	-0.438***	-0.443***
	(0.0725)	(0.0729)	(0.0725)	(0.0729)	(0.0724)	(0.0729)	(0.0724)	(0.0742)
N_worst	-0.471	0.113						
	(0.336)	(0.533)						
HR screen x N_worst		-1.201						
		(0.869)						
Skills test screen x N_worst		-0.562						
		(0.724)						
F_worst			-0.522*	-0.998*				
			(0.289)	(0.576)				
HR screen x F_worst				0.289				
				(0.928)				
Skills test screen x F_worst				0.845				
				(0.661)				
N_risky					-0.644***	-0.120		
					(0.236)	(0.383)		
HR screen x N_risky						-0.0226		
						(0.515)		
Skills test screen x N_risky						-1.346**		
						(0.546)		
F_risky							-0.539***	-0.574**
							(0.140)	(0.254)
HR screen x F_risky								-0.0683
								(0.377)
Skills test screen x F_risky								0.128
								(0.328)
Constant	4.892***	4.887***	4.893***	4.898***	4.896***	4.889***	4.906***	4.907***
	(0.0514)	(0.0515)	(0.0512)	(0.0514)	(0.0514)	(0.0517)	(0.0515)	(0.0522)
Observations	3,092	3,092	3,092	3,092	3,092	3,092	3,092	3,092
R-squared	0.014	0.015	0.015	0.015	0.016	0.018	0.017	0.018

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Figure 1. Conceptual Flow of Evaluation Procedures



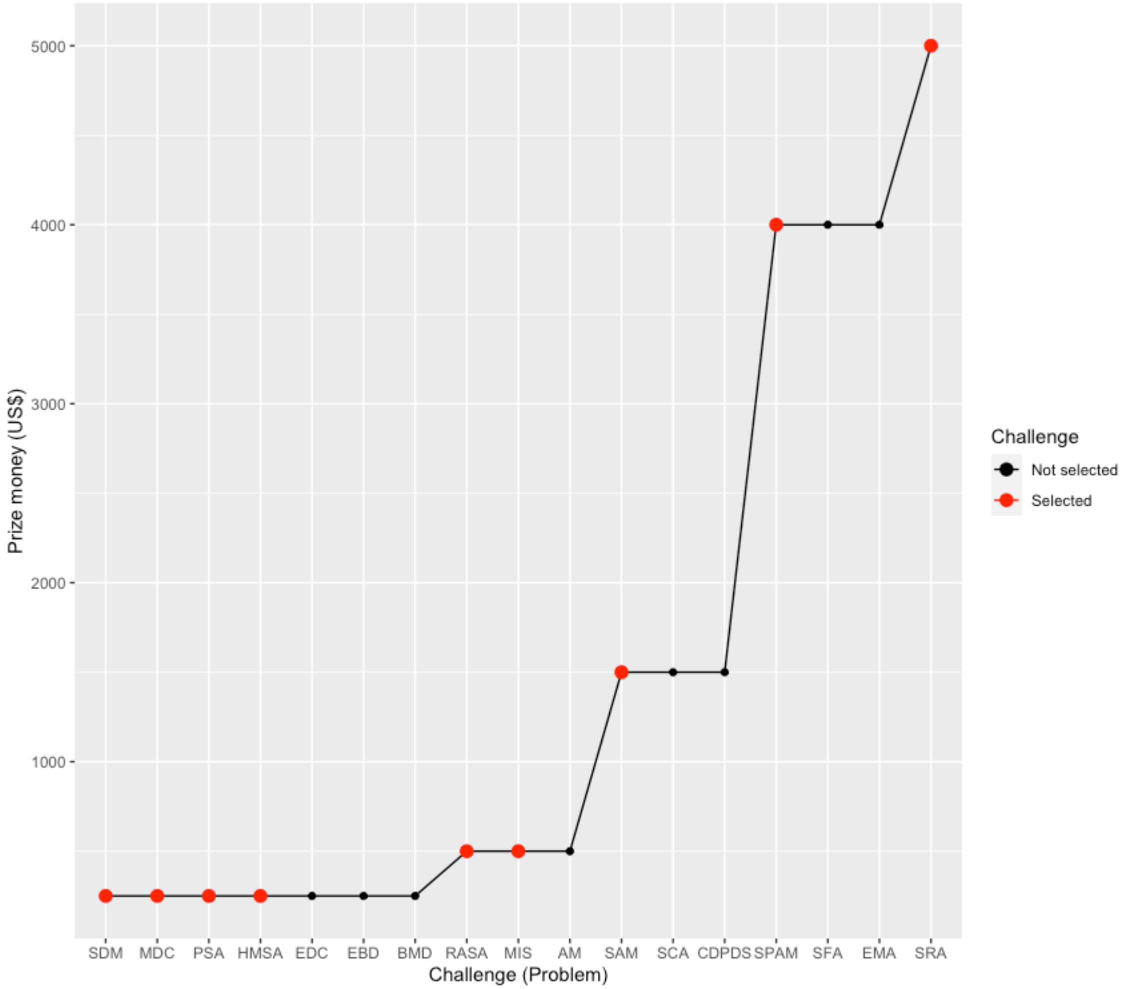
All interested individuals completed an HR screen and skills test screen

Selected three types of evaluators to evaluate 101 solutions from 9 NASA robotics challenges:
(i) Unscreened (non-expert)
(ii) HR screen (experts)
(iii) Skills test screen (experts)

Implement randomized block design:
(i) 374 evaluators each randomly assigned 2 challenges and a total of 10 solutions (5 from each challenge)
(ii) Total of 38.2 evaluations/ solution

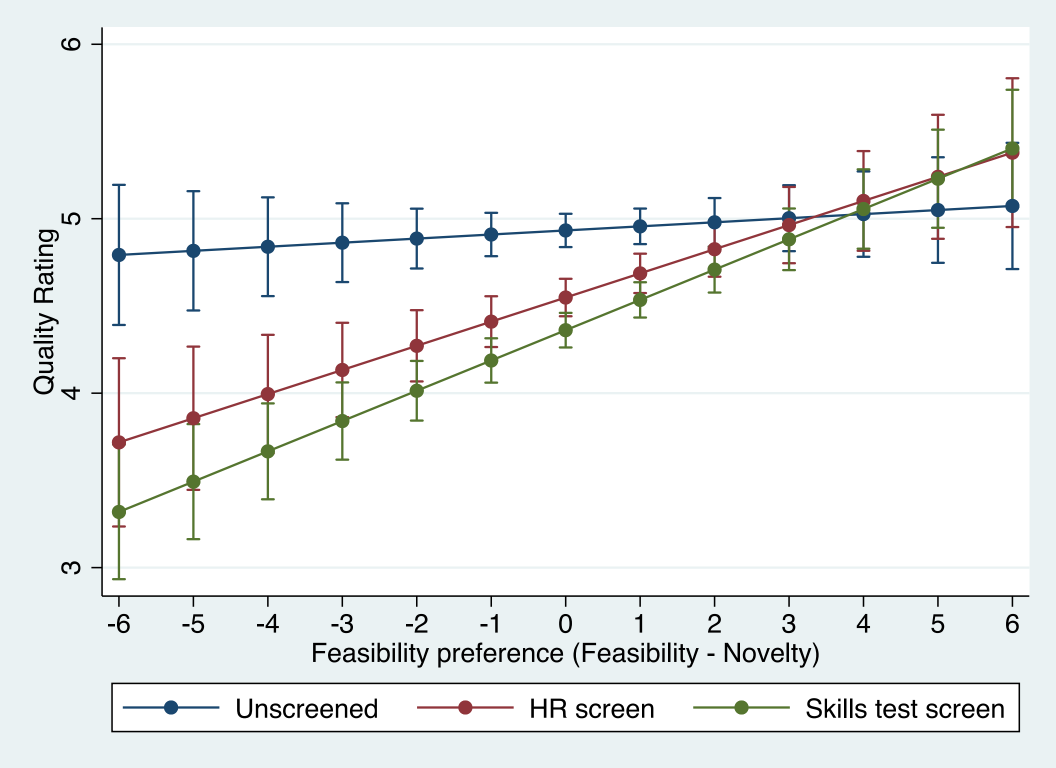
Evaluators presented with problem statement and submission guidelines, then provided numerical and narrative ratings for novelty, feasibility and quality

Figure 2. Relationship Between Challenge Prize Money and Problem Complexity



Note: We used prize money as a proxy for problem complexity (larger dollar amounts were awarded for more complex problems drawn from multiple domains, as well as the amount of time and effort estimated to come up with a solution). Based on this approach, SAM, SPAM and SRA corresponded to high complexity problems.

Figure 3. Margins Plot of Relationships between Quality Rating, Evaluator Expertise Type and Feasibility Preference



Appendix

Table A1. Summary Statistics By Challenge Block (N = 374)

	Chi-Sq/ANOVA Test
HR screen (work experience)	*p* = 0.240
Skills test screen (skills test score)	*p* = 0.449
Robotics expertise distance	*p* = 0.446
Distance to roboticist discipline	*p* = 0.803
Female	*p* = 0.965
Age range	*p* = 0.766
Bachelors	*p* = 0.019
Masters	*p* = 0.278
USA	*p* = 0.949

Note: Solutions were exogenously assigned to evaluators using a randomized block design, where each evaluator was first, randomly assigned two of nine challenges, and then randomly assigned five solutions to evaluate within each challenge, for a total of 36 blocks in the design. There are 3 observations are deleted due to missing covariate data.

Table A2. Estimated Relationships Between Quality Rating, Evaluator Expertise Type and Feasibility Preference (Categorical Variable)

	Dependent Variable: Quality Rating						
VARIABLES	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7
HR screen	-0.356***	-0.360***	-0.673***	-0.803***	-0.645***	-0.621**	-0.512***
	(0.0718)	(0.0718)	(0.151)	(0.151)	(0.152)	(0.200)	(0.162)
Skills test screen	-0.576***	-0.599***	-0.634***	-0.731***	-0.603***	-0.607***	-0.511***
	(0.0677)	(0.0679)	(0.134)	(0.134)	(0.135)	(0.0970)	(0.134)
N = F		0.149*	0.156	0.136	0.184	0.170	0.206
		(0.0762)	(0.123)	(0.118)	(0.118)	(0.146)	(0.133)
F > N		0.415***	0.0984	0.0862	0.143	0.138	0.167
		(0.0717)	(0.121)	(0.119)	(0.119)	(0.137)	(0.130)
HR screen x N = F			0.334*	0.307	0.299	0.299	0.217
			(0.191)	(0.188)	(0.186)	(0.183)	(0.190)
HR screen x F > N			0.470**	0.478***	0.437**	0.395*	0.194
			(0.185)	(0.183)	(0.184)	(0.202)	(0.197)
Skills screen x N = F			-0.307*	-0.301*	-0.270	-0.222	-0.251
			(0.178)	(0.176)	(0.176)	(0.204)	(0.174)
Skills screen x F > N			0.438***	0.462***	0.488***	0.521**	0.334**
			(0.166)	(0.165)	(0.166)	(0.191)	(0.166)
Novelty confidence				0.127***	0.143***	0.143***	0.156***
				(0.0370)	(0.0378)	(0.0413)	(0.0387)
Feasibility confidence				0.0926**	0.102***	0.110**	0.0826**
				(0.0377)	(0.0382)	(0.0391)	(0.0372)
Constant	4.967***	4.760***	4.862***	3.712***	4.206***	4.136***	4.169***
	(0.0471)	(0.0697)	(0.0970)	(0.161)	(0.223)	(0.502)	(0.270)
Evaluator controls	N	N	N	N	Y	Y	Y
Challenge FE	N	N	N	N	N	Y	N
Solution FE	N	N	N	N	N	N	Y
Observations	3,869	3,869	3,869	3,869	3,830	3,830	3,830
R-squared	0.018	0.026	0.034	0.059	0.092	0.094	0.095

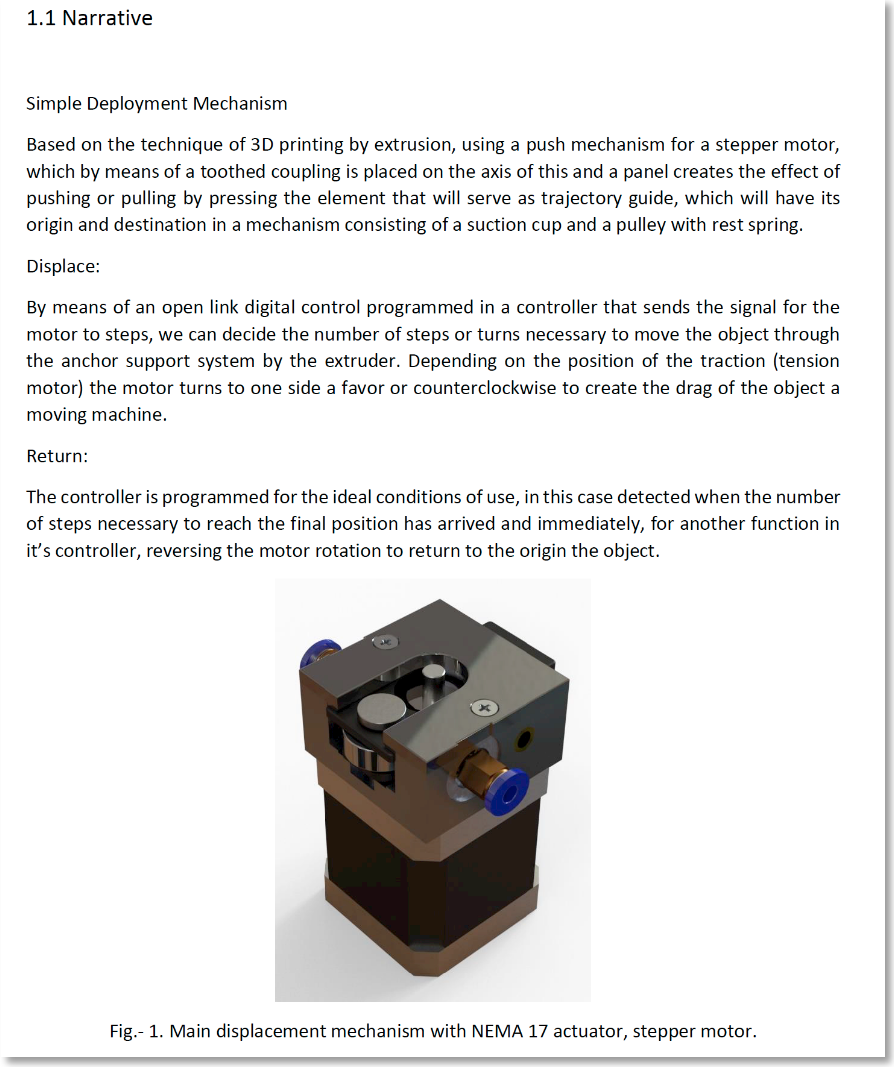
Note: N = F means that the evaluator gave the same novelty and feasibility rating to the solution; F > N means that the evaluator gave a higher feasibility than novelty rating; N > F is the baseline category and means that the evaluator gave a higher novelty than feasibility rating. Robust standard errors in parentheses; *** p<0.01, ** p<0.05, * p<0.1

Table A3. Estimated Relationships between Quality Rating, Evaluator Expertise Type, Feasibility Preference and Problem Complexity

	DV: Quality Rating					
VARIABLES	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
HR screen	-0.359***	-0.413***	-0.550***	-0.411***	-0.389**	-0.383***
	(0.0717)	(0.0913)	(0.0886)	(0.0903)	(0.126)	(0.0770)
Skills test screen	-0.593***	-0.643***	-0.736***	-0.608***	-0.566***	-0.541***
	(0.0676)	(0.0869)	(0.0844)	(0.0861)	(0.136)	(0.103)
Feasibility preference	0.106***	0.0483	0.0350	0.0417	0.0474	0.0714
	(0.0187)	(0.0384)	(0.0374)	(0.0377)	(0.0570)	(0.0431)
High complexity (HC)	0.109*	0.0449	0.0289	-0.0415		
	(0.0606)	(0.101)	(0.0983)	(0.0978)		
HR screen x Feas. pref.		0.0866	0.0992*	0.0826	0.0639	0.0180
		(0.0587)	(0.0577)	(0.0580)	(0.0448)	(0.0559)
Skills screen x Feas. pref.		0.0488	0.0655	0.0695	0.0655	0.0194
		(0.0532)	(0.0521)	(0.0527)	(0.0706)	(0.0527)
HR screen x High complexity		0.0725	0.0676	0.0850	0.0672	0.0943
		(0.155)	(0.153)	(0.151)	(0.192)	(0.130)
Skills screen x High complexity		0.0419	0.0588	0.124	0.0901	0.0987
		(0.145)	(0.144)	(0.143)	(0.146)	(0.151)
High complexity x Feas pref.		-0.112*	-0.0878	-0.0588	-0.0639	-0.103
		(0.0650)	(0.0651)	(0.0685)	(0.0883)	(0.0727)
HR screen x HC x Feas. pref.		0.139	0.111	0.100	0.111	0.102
		(0.103)	(0.103)	(0.107)	(0.0925)	(0.112)
Skills screen x HC x Feas. pref.		0.313***	0.284***	0.263***	0.273**	0.255***
		(0.0862)	(0.0866)	(0.0895)	(0.107)	(0.0835)
Novelty confidence			0.120***	0.138***	0.138**	0.153***
			(0.0373)	(0.0381)	(0.0456)	(0.0393)
Feasibility confidence			0.0957**	0.104***	0.112**	0.0849**
			(0.0378)	(0.0384)	(0.0380)	(0.0376)
Constant	4.902***	4.943***	3.806***	4.375***	4.279***	4.328***
	(0.0518)	(0.0600)	(0.140)	(0.206)	(0.462)	(0.243)
Evaluator controls	N	N	N	Y	Y	Y
Challenge FE	N	N	N	N	Y	N
Solution FE	N	N	N	N	N	Y
Observations	3,869	3,869	3,869	3,830	3,830	3,830
R-squared	0.026	0.033	0.057	0.090	0.092	0.093

Robust standard errors in parentheses; *** p<0.01, ** p<0.05, * p<0.1

Figure A1. Sample Solution Narrative



1.1 Narrative

Simple Deployment Mechanism

Based on the technique of 3D printing by extrusion, using a push mechanism for a stepper motor, which by means of a toothed coupling is placed on the axis of this and a panel creates the effect of pushing or pulling by pressing the element that will serve as trajectory guide, which will have its origin and destination in a mechanism consisting of a suction cup and a pulley with rest spring.

Displace:

By means of an open link digital control programmed in a controller that sends the signal for the motor to steps, we can decide the number of steps or turns necessary to move the object through the anchor support system by the extruder. Depending on the position of the traction (tension motor) the motor turns to one side a favor or counterclockwise to create the drag of the object a moving machine.

Return:

The controller is programmed for the ideal conditions of use, in this case detected when the number of steps necessary to reach the final position has arrived and immediately, for another function in it's controller, reversing the motor rotation to return to the origin the object.

Fig.- 1. Main displacement mechanism with NEMA 17 actuator, stepper motor.

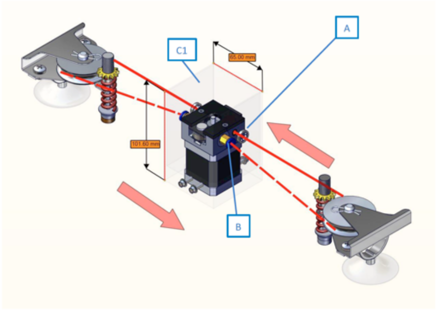
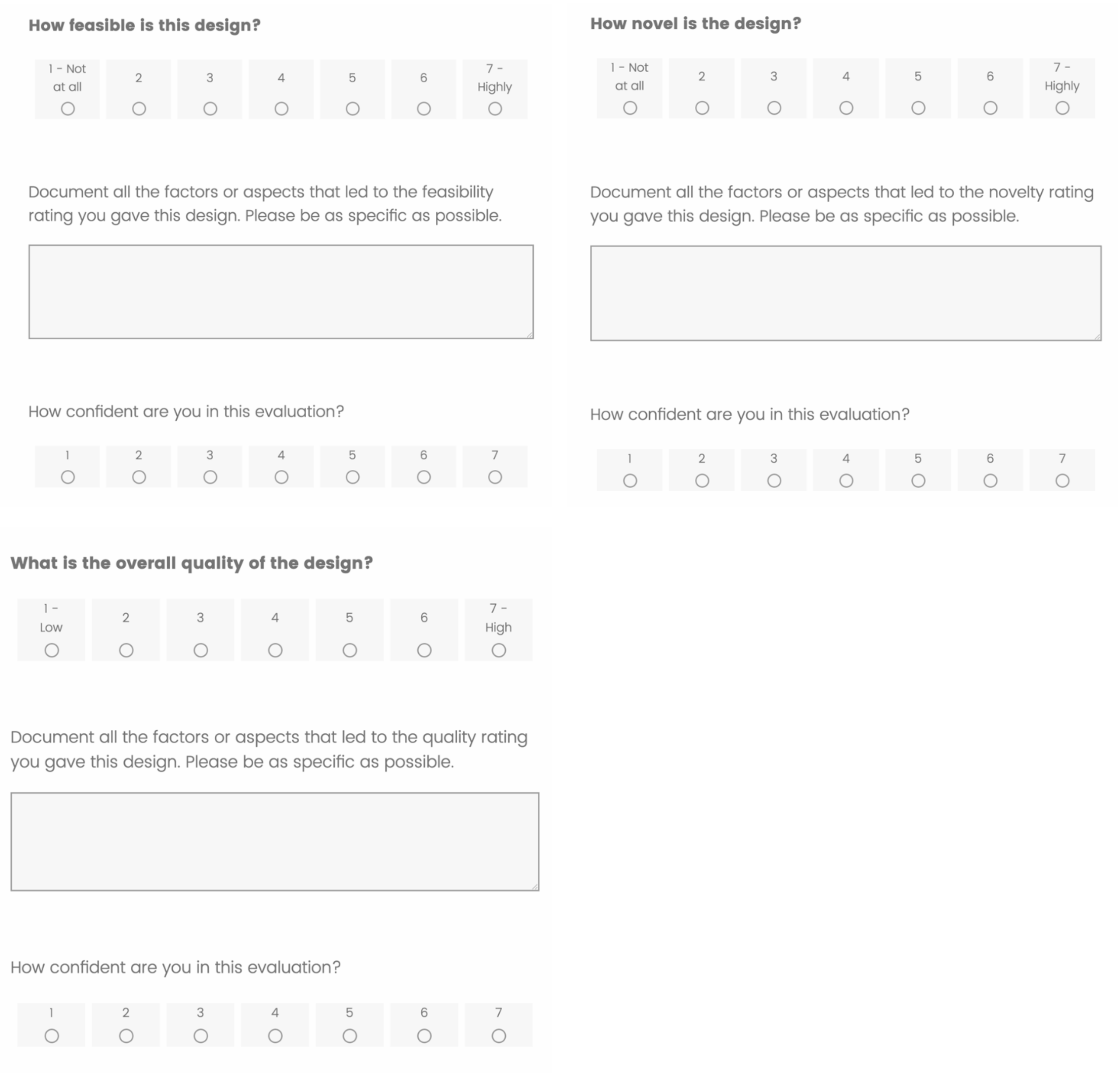


Fig.- 2. Simple deployment mechanism with spring system of suction cups and a pulley with rest spring.

Figure A2. Screenshots of Evaluation Procedures



How feasible is this design?

1 - Not at all	2	3	4	5	6	7 - Highly
○	○	○	○	○	○	○

Document all the factors or aspects that led to the feasibility rating you gave this design. Please be as specific as possible.

How confident are you in this evaluation?

1	2	3	4	5	6	7
○	○	○	○	○	○	○

How novel is the design?

1 - Not at all	2	3	4	5	6	7 - Highly
○	○	○	○	○	○	○

Document all the factors or aspects that led to the novelty rating you gave this design. Please be as specific as possible.

How confident are you in this evaluation?

1	2	3	4	5	6	7
○	○	○	○	○	○	○

What is the overall quality of the design?

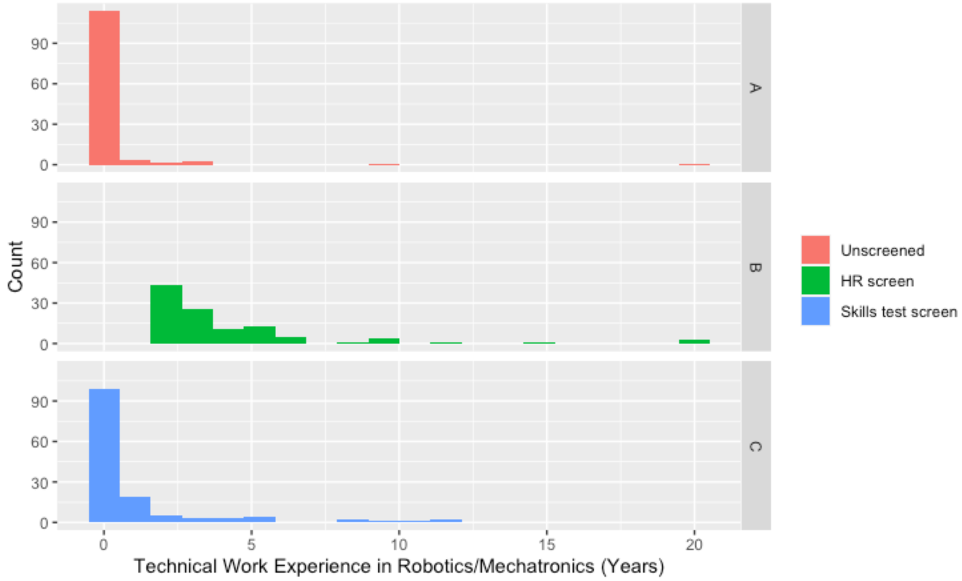
1 - Low	2	3	4	5	6	7 - High
○	○	○	○	○	○	○

Document all the factors or aspects that led to the quality rating you gave this design. Please be as specific as possible.

How confident are you in this evaluation?

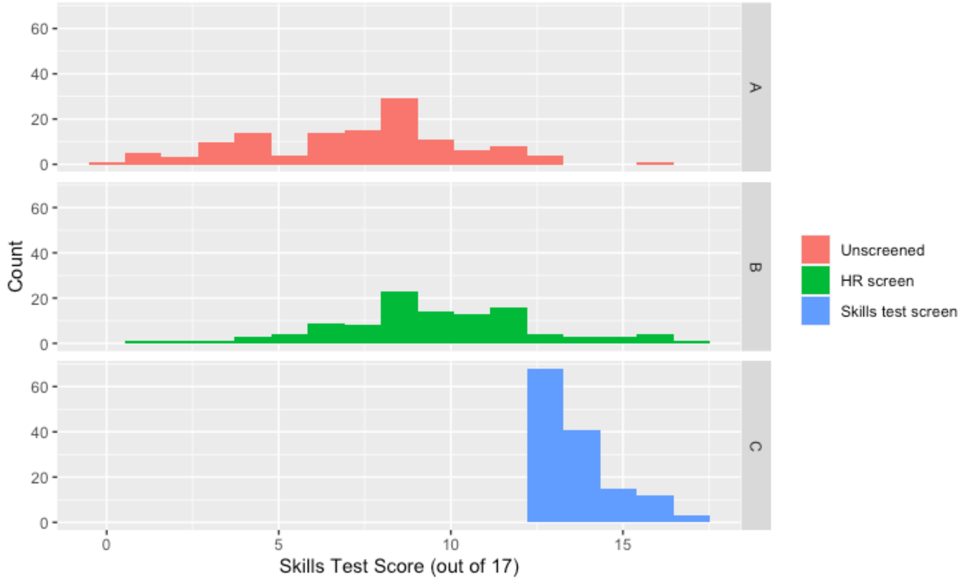
1	2	3	4	5	6	7
○	○	○	○	○	○	○

Figure A3. Distribution of Work Experience in Robotics/Mechatronics Engineering (N = 374)



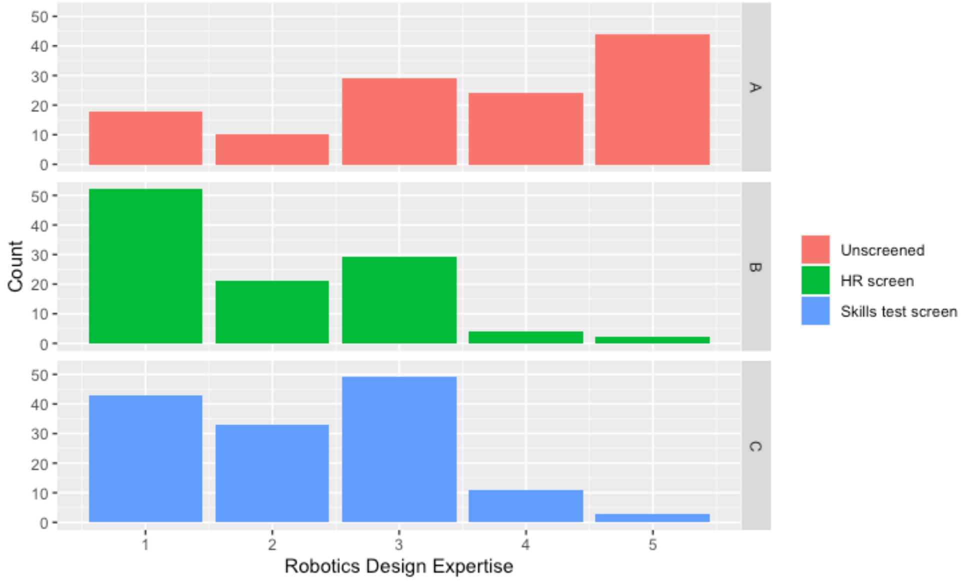
Note: Two years or more of robotics work experience is required for HR screen threshold.

Figure A4. Distribution of Skills Test Score (N = 374)



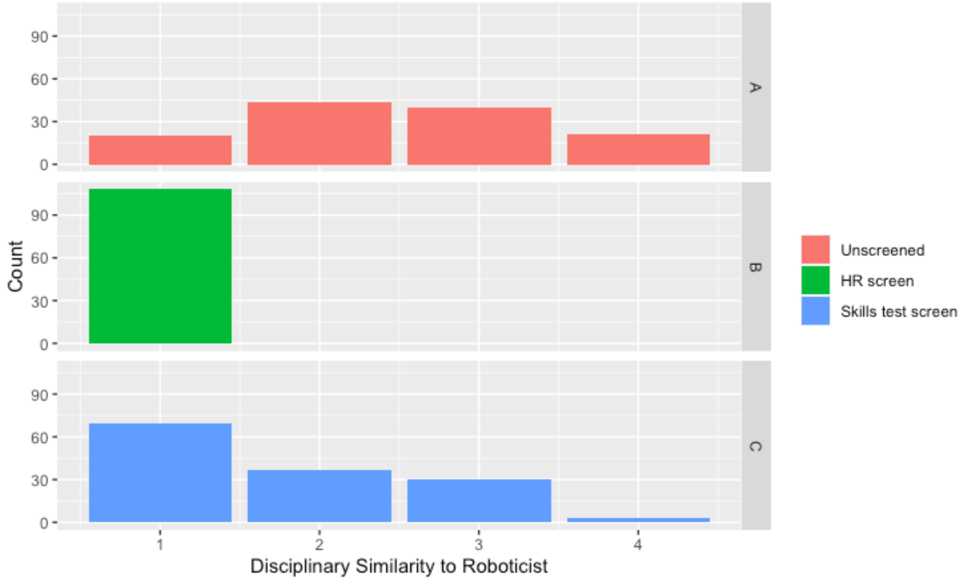
Note: 13 or more out of 17 is required for skills test screen threshold.

Figure A5. Distribution of Robotics Design Expertise



Note: The Robotics Design Expertise measure is scored on a five-point Likert scale, where 1 = inside my field of expertise, 3 = at the boundary of field of expertise, 5 = outside my field of expertise.

Figure A6. Distribution of Disciplinary Similarity to Roboticist



Note: The categories for the Disciplinary Similarity to a Roboticist measure is based on disciplinary classifications in Szjanfarber et al., 2020, where 1 = most similar to a roboticist, 4 = least similar to a roboticist.

Figure A7. Margins Plot with 95% CIs of Relationships between Quality Rating, Evaluator Expertise Type and Novelty Rating

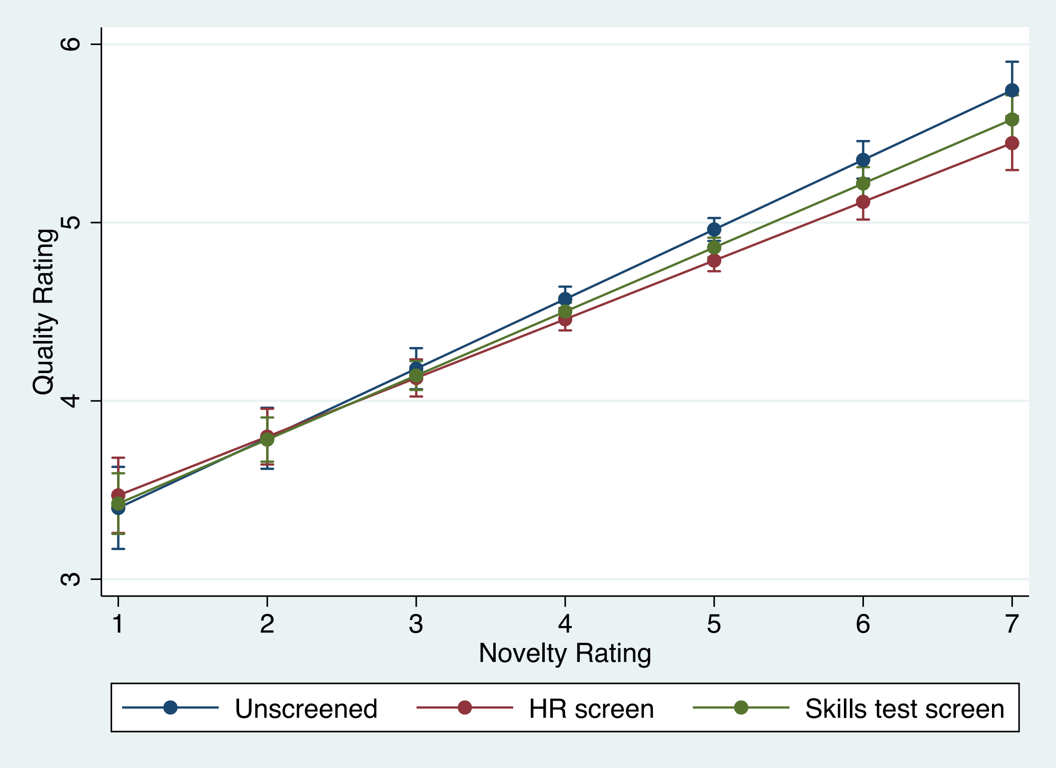


Figure A8. Margins Plot with 95% CIs of Relationships between Quality Rating, Evaluator Expertise Type and Feasibility Rating

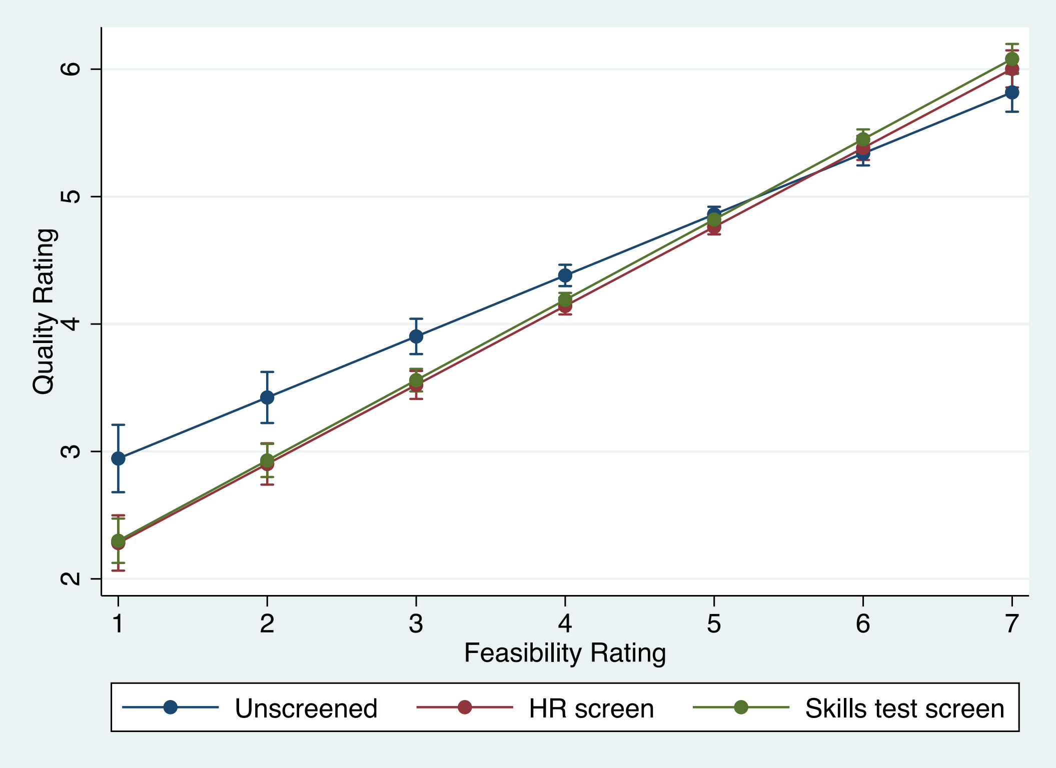
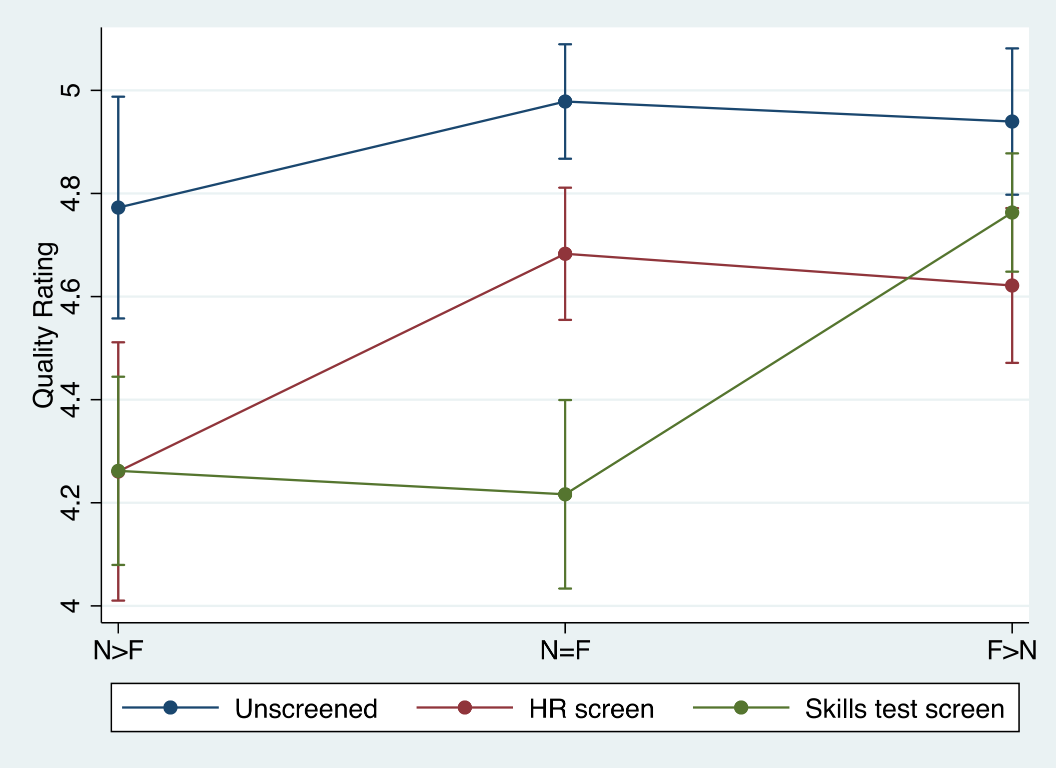


Figure A9. Margins Plot with 95% CIs of Relationships Between Quality Rating, Evaluator Expertise Type and Feasibility Preference (Categorical Variable)



Note: Margins plots based on regression coefficients from Table A2 Model 3.