When Does Project Feasibility Drive Technological Innovation?

Evaluator Expertise Range, Architectural Knowledge, and Preferences for Existing Technologies

Jacqueline N. Lane[1,3*], Zoe Szajnfarber[2,3], Jason Crusan[2,3], Michael Menietti[1,3] and Karim R. Lakhani[1,3]

[1]Harvard Business School
[2]George Washington University
[3]The Digital, Data, and Design (D^3) Institute at Harvard

[*]Correspondence to: jnlane@hbs.edu

This version: January 10, 2023

Abstract

The creation of technological innovations draws on knowledge of both the components that comprise the system and architectural knowledge of how the components are interconnected into a holistic system. Often, this knowledge integrates multiple domains of expertise. Yet we know relatively little about how expertise range—or the degree of knowledge overlap between an evaluator's domain(s) of expertise and the knowledge domains embodied in the problem area—shapes an evaluator's perceptions of the design and their effects on preferences for more novel versus more feasible solutions. We partnered with NASA and Freelancer.com to design an evaluation challenge, recruiting 374 evaluators from inside and outside the domains of aerospace and robotics design, to rate 101 unique solutions for a total of 3,850 evaluator-solution pairs. Our results show that univalent evaluators—with expertise in a single domain of the problem area—are more likely to prefer more feasible and less novel solutions, demonstrating a feasibility preference. This feasibility preference is attenuated when evaluators have multivalent expertise spanning both domains of the problem area. Topic modeling of the evaluators' open-text comments suggests that whereas univalent evaluators assess the details of a design's components, multivalent evaluators focus more holistically on the design's architecture and functionality. Our findings suggest that selecting evaluators with varied expertise range in the problem area can lead to greater resource allocation to novel, risky ideas that favor less familiar, untested technologies.

Keywords: evaluations, expertise, feasibility preference, technological innovation, architectural knowledge, resource allocation, field experiment

1. Introduction

A firm's technological trajectory hinges on its ability to generate new ideas and select among potential alternatives that have the capacity to increase performance (Berg 2016, Boudreau et al. 2016, Criscuolo et al. 2017, Levinthal 1997). While much scholarly attention has focused on how firms can improve their idea generation and problem-solving capabilities, such as through ideation contests (Jeppesen and Lakhani 2010), crowdsourcing (Dahlander et al. 2019, Mollick 2014), social interactions (Catalini 2018, Hasan and Koning 2019, Lane, Ganguli, et al. 2021), and monetary incentives (Myers 2020), to cast a wider net for potential solutions (Dahlander et al. 2019), we know relatively less about how firms should select evaluators to determine which ideas within their research and development (R&D) pipelines should be earmarked for further consideration and implementation (Criscuolo et al. 2017). This disconnect exists even though many firms, scientific institutions, and government agencies, alike, spend significant time and attention on evaluation and selection processes (Bower 1972, Lamont 2009, Maritan and Lee 2017, Noda and Bower 1996), often sourcing experts with deep knowledge of the problem area to assess alternative projects for their quality (Azoulay et al. 2019, Bian et al. 2021, Criscuolo et al. 2017, Lane et al. 2021, Lee 2015). Perhaps a critical factor that makes evaluation and selection processes challenging for technological innovations, such as medical imaging, driverless autonomous vehicles, facial recognition, and humanoid robots, is that effective problem-solving and evaluation of alternative solutions often require knowledge of multiple domains that are integrated to develop a complete product or innovation. In this paper, we investigate how an evaluator's *range* of domain-relevant knowledge in the problem area—or the degree of overlap between an evaluator's domain(s) of expertise and the knowledge domains of the problem area—systematically shapes their perceptions of a technological design.

The importance of expertise range in the problem area on evaluation outcomes stems from the notion that developing new technological innovations involves two types of knowledge. The first type is component knowledge about the core design concepts or technologies used in the components of a design, and the second type is architectural knowledge of how the components are linked together into a holistic system (Henderson & Clark, 1990). We propose that an evaluator's range of expertise in the relevant domains of the problem area is likely to shape their perceptions of a design either from a detailed view, as the collection of its parts or components that comprise the system, or from a holistic view, as a system. This distinction between these two types of knowledge can be better understood by drawing an analogy to the technology behind fully autonomous self-driving cars—defined as vehicles that can operate themselves and perform necessary functions without any human attention or intervention (Davenport & Desmond, n.d.). Autonomous driving requires three main components: sensors, semiconductors, and software (Coffin et al., 2019). Whereas sensors enable vehicles to see road conditions at various distances, semiconductors facilitate the processing of data gathered by sensors to make real-time driving decisions, which are backed by machine learning and mapping software to support frequent updates (Coffin et al., 2019). The notion that the technology behind each of these components requires differentiated domains of expertise is exemplified by the different players that are investing in the technology behind autonomous vehicles. While automakers, such as Ford and General Motors, and Tesla are making significant investments to develop and manufacture self-driving vehicles—i.e., the holistic system, many of the technological capabilities are being developed in partnership with technology companies with the domain expertise to produce the sensors, semiconductors, and software—i.e., the components being used to operate autonomous vehicles. For instance, Microsoft has partnered with Volkswagen to develop a cloud-based system that will use Microsoft Azure for computation, artificial intelligence (AI), and storage, while chip designer, NVIDIA, which already has several automotive customers, is developing the hardware that will unify a wide range of in-car technology in a single platform (Gray, 2022). Although efforts in the sector remain laser-focused on establishing the feasibility of fully autonomous vehicles, as attention turns to evaluating alternative self-driving car options, evaluators may have differentiated assessments of the design, depending on whether they perceive it as a system or its different components and functions.

We relate expertise range in the problem area, or degree of domain overlap with the domain(s) of the problem to be solved, to evaluators' perceptions of a design either by its components or as an entire, holistic system. Domain expertise refers to familiarity with the knowledge, technical skills, techniques, equipment, opinions, paradigms, concepts, and aesthetic criteria of the focal domain (Amabile 1983, Csikszentmihalyi 1990, Dane 2010, Hinds et al. 2001, Shanteau 1992). As most engineering problems tend to draw on knowledge from two or more domains, potential evaluators may possess varying degrees of overlap between their domain knowledge and the domains corresponding to the problem area. Whereas individuals with *univalent* expertise in a single domain of the problem area will be more likely to perceive the details of a design, as a compilation of different parts or components—due to the partial overlap between their domain expertise and the knowledge embodied in the problem area, individuals with *multivalent* expertise in the problem area will be more likely to perceive the design as a holistic system—resulting from their broader or more complete knowledge overlap between their domains of expertise and the knowledge domains embodied in the problem area. The objective of this paper is to determine the extent that these alternative perspectives on the design change evaluators' perceptions of a design's quality—where quality is defined here as a novel solution that can be feasibly implemented to solve the problem at hand (Amabile, 1983; Onarheim, 2012; Rindova & Petkova, 2007). Although this definition suggests that high-quality designs achieve a balance of novelty and feasibility, it is possible that evaluators' expertise range in the problem area—univalent or multivalent—may alter their preferences toward more feasible or novel solutions when assessing a given design for its quality. Given the importance of designing new and feasible solutions to technological problems, it is important to consider the extent to which the range of an evaluator's expertise may shape their preferences for solutions that are more balanced in terms of their perceived novelty and feasibility.

To systematically investigate the relationships between evaluator expertise range and evaluation outcomes requires a rich dataset that directly observes how expertise in the problem area relates to evaluators' subjective assessments of technological designs in terms of their novelty, feasibility, and overall quality. To draw causal inferences between evaluator expertise range and evaluation outcomes, we designed and executed a field experiment with the U.S. National Aeronautics and Space Administration (NASA) and Freelancer.com, an online labor marketplace. We collaborated with NASA to host a space robotics evaluation challenge series and worked with the agency to recruit evaluators from the Freelancer.com community using an open "broadcast search" for registrants (Jeppesen and Lakhani 2010). We chose this setting for two reasons. The first is due to the centrality of evaluation and selection processes on resource allocation decisions within NASA. The second is the importance of different domains of expertise in the problem area—i.e., univalent expertise in robotics *or* aerospace and multivalent expertise in robotics *and* aerospace—that may shape how evaluators perceive technological designs as either the sum of its components or as an entire system, respectively—as well as the implications of these perceptions on evaluators' preferences for more feasible or more novel solutions in evaluations of design quality.

To sample from a diversity of evaluator expertise in the problem area, we collected information on the registrants' background and experience, as well as their technical skills in robotics by administering a robotics skills test assessment. Overall, we mobilized a total of 374 unique evaluators that were randomly drawn from both inside and outside the problem area of aerospace robotics design, and randomly assigned each evaluator to judge a subset of the 101 designs, for a total of 3,850 evaluator-solution pairs across nine challenges. The evaluators rated multiple solutions, judging each solution for its novelty, feasibility, and overall quality, and provided open-text comments to justify their choices. This approach enabled us to examine how univalent versus multivalent expertise in the problem area (versus no expertise in the problem area) informed evaluators' preferences for feasibility and novelty in problem solutions. It also allowed us to correct for idiosyncrasies associated with a single evaluator or solution (Amabile & Mueller, 2008), and to make inferences based on perceptions of quality without observing true quality, due to the difficulty in observing ground truth in technology-driven innovations (Boudreau et al. 2016).

We report several noteworthy patterns. First, our experimental results show that, compared to non-expert ("outside") evaluators, evaluators with either univalent or multivalent expertise provide harsher (i.e., more critical) judgments of design feasibility; univalent and multivalent experts do not differ in the comparisons to non-experts. Second, we observe that univalent evaluators exhibit a *feasibility preference*, in which evaluators with knowledge overlap in a single domain of the problem area (either robotics or aerospace, but not both), are more likely to perceive solutions as higher in quality when they are higher in feasibility and lower in novelty. Indeed, relative to the non-expert outside evaluators, those with univalent expertise weighed the feasibility of a design by nearly 30 percent more in their evaluations of the design's quality; outside evaluators exhibited comparatively more balanced weightings of novelty and feasibility.

Third, our main finding indicates that the feasibility preference is attenuated in evaluators possessing multivalent expertise spanning both domains of the problem area. That is, multivalent evaluators are more likely than univalent evaluators to perceive that a high-quality design is a novel solution that can be feasibly implemented into a workable device. Lastly, we perform natural language processing of the evaluators' open-text comments to gain deeper insights into the observed patterns. Topic modeling and sentiment analyses of the evaluators' comments indicate that expertise range alters how evaluators perceive a design's features. Whereas univalent expertise enables evaluators to see more *details* about a design's components, allowing them to uncover more weaknesses in a design's feasibility relative to outside evaluators, multivalent expertise facilitates more *holistic* assessments of a design's architecture—i.e., the interconnections between its components, facilitating assessments of the design's functionality. This multivalent expertise enables these evaluators to recognize and appreciate the potential upsides associated with more novel and feasible designs.

Our findings have practical implications for innovative firms incentivized to develop new technologies that disrupt existing organizational routines or capabilities (Dewar & Dutton, 1986; Ettlie et al., 1984; Henderson & Clark, 1990; Tushman & Anderson, 1986). Although evaluation processes will benefit from drawing upon the expertise of component-oriented univalent evaluators and the systems-oriented multivalent evaluators, it is the multivalent evaluators who possess the relevant alternative perspectives (Hong & Page, 2001; Jeppesen & Lakhani, 2010) to simultaneously recognize and value more novel yet feasible solutions. This suggests that the ability to develop more significant changes to a firm's existing products or technologies depends in part, on bringing together an appropriate collection of evaluators who hold architectural knowledge of each relevant domain of the problem area.

2. The Development and Evaluation of Technological Innovations in Engineering

Engineering design, which is the province of our study, is best understood as a problem-solving activity, in which designers use knowledge to design, produce, and operate artifacts with concrete objectives (Onarheim, 2012; Stacey & Eckert, 2010; Vincenti, 1990). Often, engineering problems are *constrained* by various pre-specifications that need to be satisfied in the design or solution. These may include performance requirements, such as power, weight, fuel consumption, and acceleration needs, as well as operational capabilities and regulations requirements, such as safety, security, and regulatory standards (Acar et al., 2019; Shah et al., 2003). Most engineering designs begin by following a well-established set of techniques acquired through formal training in the domain, prioritizing a dominant design with a stable architecture

(Henderson & Clark, 1990). The starting point for solving many engineering problems begins with reconfiguring existing, “off-the-shelf” technologies that have been previously cataloged and recorded in the form of reference works in textbooks, handbooks, professional journals, company reports, as well as government publications (Florman 1987, Vincenti 1990). Due to a focus on reliability and consistency, the objective of engineering design for many established technology firms is to limit the amount of novelty in a product or innovation in terms of new components or new uses of existing components in favor of more incremental approaches that are more likely to be feasible (Dewar & Dutton, 1986; Henderson & Clark, 1990; Stacey & Eckert, 2010; Tushman & Anderson, 1986).

The standard approach to engineering design raises the question of when novel technologies might arise. The downside of leveraging existing technologies to accomplish current design goals is that new designs may be confined by the parts of an earlier design that need to be retained (Stacey & Eckert, 2010). This may be particularly problematic when the problem’s pre-specifications or constraints interact with one another in complex ways (i.e., beyond pairwise interactions) and give rise to conflicting demands (Stacey & Eckert, 2010). As it may not be possible to meet all the problem’s pre-specifications, a design may forgo or postpone certain pre-specifications to reduce complexity or uncertainty (Onarheim, 2012). Weaker pre-specifications may be relaxed to allow for less ideal but feasible designs, while stronger pre-specifications must be met (Stacey & Eckert, 2010). It is during this design phase of prioritizing, adding, and dropping pre-specifications that novelty—in the form of new components or parts may emerge because the use of existing parts or components in a design cannot satisfy the needs of the problem or task (Florman 1987, Layton Jr 1974, Vincenti 1990). Thus, even though novelty may not a requirement of a solution from the outset (Layton Jr 1974), it occupies a fundamental role in engineering design. Novel designs expand the design space—or the count of all possible solutions for a given problem—and they may even lead to the creation of new linkages and interconnections with the other components that may change the architecture of a design (Henderson & Clark, 1990). Thus, novelty pushes the limits of the design space (Newell 1979, Newell and Simon 1972) that may advance engineering knowledge (Vincenti 1990) and the technological frontier. Yet, increasing the novelty of a design may create a potential tradeoff with its feasibility, as novel technology and approaches are less likely to be tested, validated, or established (Boudreau et al., 2016; Thomke, 1998).

The evaluation of alternative solutions or designs is critical to assessing their quality, particularly as engineering problems tend to have many pre-specifications that need to be satisfied. The design literature suggests that an engineering design not only needs to be novel (i.e., original or unexpected) but it must also have practical utility and satisfy some intended functions to desired specifications. In other words, engineering design quality is judged by the extent to which it is novel and can be feasibly implemented to solve the problem at hand (Amabile, 1983; Onarheim, 2012; Rindova & Petkova, 2007). An evaluator’s range of expertise in the problem area of the engineering design task is likely to shape their assessments of a design's novelty and feasibility, as well as its overall quality.

Figure 1 depicts the different kinds of knowledge overlap in our focal problem area of aerospace robotics. In our study context, evaluators may possess outside expertise, univalent expertise in either aerospace or robotics, or multivalent expertise in both aerospace and robotics. Whereas univalent evaluators' knowledge of a single domain in the problem area may facilitate a higher understanding and familiarity of specific components in the design relative to outside expertise evaluators, multivalent evaluators' range of expertise across both domains in the problem area may facilitate an understanding of not only the design's components but also the interconnections or linkages between them, offering them a systems-level view of the problem and the proposed solutions. In the remainder of the section, we theorize the effects of expertise range on evaluators' perceptions of feasibility, as well as their preferences for more feasible or novel designs when judging the overall quality of alternative solutions.

[Figure 1 about here]

2.1. Relationships Between Evaluator Expertise Range and Perceptions of Design Feasibility

The ability to design safe, reliable, and reproducible artifacts is an inherent need of all engineering innovations (Chai et al. 2021, Vincenti 1990). As engineering quality is the ability to create novel solutions that can be feasibly implemented (Amabile, 1983; Shah et al., 2003), this essential need for feasibility is reinforced by the potentially catastrophic consequences of failure that can outweigh the potential upsides of novelty (Chai et al., 2021). Consequently, a fundamental requirement of the evaluation process is to identify feasible designs that satisfy the problem's critical pre-specifications and can be implemented into a workable device with relative ease (Shah et al. 2003). As feasibility is a concrete requirement that is based on relevant facts and evidence, as opposed to personal feelings, attitudes, emotions, or beliefs (Ferioli et al. 2010), domain-relevant expertise in the problem area may lead to more critical judgments of a design's feasibility.

Domain expertise has two relevant knowledge components during evaluations of engineering designs: analytical and experiential knowledge (Shah et al., 2003). Whereas analytical knowledge entails a procedural understanding of the formulas, techniques, and theoretical concepts being applied in the design, experiential knowledge entails an understanding of prior designs, as well as "rules of thumb" or heuristics—such as rough checks, shortcuts, or calculations (Chi et al. 1981, Newell and Simon 1972, Simon 1978) that enable individuals to detect limitations in a design, including fatal flaws (Camerer & Johnson, 1991). Hence, evaluators whose expertise falls outside of the problem area will be limited in their judgments of a design's feasibility due to their lack of analytical or experimental knowledge in one or more of the pertinent domains of the problem area. When evaluating the choice of a design concept applied to a particular component of a design, outside expertise evaluators may not have sufficient domain knowledge—i.e., familiarity with the technical skills, techniques, equipment, opinions, paradigms, etc. (Amabile, 1983; Csikszentmihalyi, 1997; Dane, 2010; Shanteau, 1992)—to judge the feasibility of a specific component that the designer or solver has selected, out of the possible set of design concepts that can be used in the design (Clark, 1985). In contrast, univalent evaluators—who possess knowledge of a single domain in the problem area—have a deeper understanding of the design concepts available to perform a well-defined function within a design. Given their knowledge and skills, univalent evaluators may pay comparatively closer attention to the specific component(s) of a design to which their domain expertise can be more readily applied. This domain-relevant knowledge may facilitate more comprehensive assessments of the extent to which a specific component in the actual design is justified over other potential choices (Baldwin & Clark, 2000; Clark, 1985). Compared to outside expertise evaluators, univalent evaluators' knowledge in the domain will lead them to uncover more potential weaknesses in a design's components that ultimately reduce its feasibility. Thus, we hypothesize:

H1. Evaluators with univalent expertise in one domain of the problem area assign more critical (lower) feasibility scores to technological designs than outside expertise evaluators.

There are two potential yet countervailing effects of multivalent expertise on judgments of a design's feasibility, leading to two alternative hypotheses about multivalent experts' feasibility judgments relative to those of univalent experts. On one hand, if univalent expertise improves the ability of evaluators to detect more problems in the feasibility of a design's components, then we might expect that multivalent expertise may lead to a more thorough uncovering of a design's limitations. Multivalent evaluators have a strong range of knowledge overlap between their domain expertise and the knowledge represented in the problem area. Due to their extensive range of domain-relevant expertise, multivalent evaluators can closely attend to each component of a design, scrutinizing the mechanism behind each component more carefully and with greater precision, which may facilitate the identification of a greater number of weaknesses in a design's feasibility. Thus, a broader expertise range may facilitate more stringent evaluations of a design's feasibility, due to multivalent evaluators' ability to detect more problems, concerns, and weaknesses that limit a design's implementation into a workable device.

On the other hand, if multivalent evaluators are instead viewing the design as a system, as opposed to the sum of its components, then their architectural knowledge of a design's interconnections (Henderson & Clark, 1990), may result in a more holistic perception of a design's feasibility. Multivalent evaluators may view that the design needs to satisfy an alternative set of requirements or pre-specifications. Rather than scrutinizing each component in more detail, multivalent evaluators may instead, be more lenient about the design concepts (Clark 1985) underlying the individual components selected for the design. They may view certain constraints as less critical to the task at hand (Stacey & Eckert, 2010), or approach the problem at a more abstract level—removing certain details and implicit assumptions (Eckert et al., 2012)—as long as the interrelationships between the components improve the overall functionality and feasibility of the design. Thus, we hypothesize two alternative hypotheses:

Hypothesis H2a. Evaluators with multivalent expertise spanning all domains of the problem area assign more critical (lower) feasibility scores to technological designs than univalent evaluators.

Hypothesis H2b. Evaluators with multivalent expertise spanning all domains of the problem area assign less critical (higher) feasibility scores to technological designs than univalent evaluators.

2.2. Evaluator Expertise Range and Quality Judgments Among Alternative Designs

A design's overall quality is determined by the extent to which it is novel and can be feasibly implemented into a workable product or innovation (Amabile, 1983; Onarheim, 2012; Rindova & Petkova, 2007). In assessments of overall quality, novelty and feasibility correspond to distinct epistemic criteria, and evaluators often need to interpret (often implicitly) how to weigh a design's novelty and feasibility in their evaluations of overall quality (Lee, 2015). These decisions may differ across evaluators (Lamont 2009, Langfeldt 2004), perhaps as a function of their expertise range in the problem area. In this section, we investigate the relationship between an evaluator's range of expertise in the problem area—univalent or multivalent—and the tendency to prioritize a design's perceived feasibility over its novelty when judging its overall quality.

As achieving a feasible approach tends to be an essential component of an engineering design's quality (Dean et al. 2006), univalent and multivalent evaluators may view finding a feasible solution among alternatives as the focal objective of the evaluation process. This perspective may be justified, particularly if a greater range in the problem area leads to the uncovering of more limitations or weaknesses in a design's feasibility. The importance of focusing on feasible designs may be even greater in reliability-critical domains, such as aerospace, automotive, and healthcare, where novelty can be unappealing because it inserts uncertainty into the long-term performance of an innovation (Miron-Spektor & Erez, 2017; Mueller et al., 2012; Stacey & Eckert, 2010), and may entail risk of product failure (Eckert et al., 2012). Moreover, this uncertainty typically cannot be completely resolved without experimentation, prototyping, and testing (Burns & Stalker, 1961; Cannon & Edmondson, 2005; Henderson & Clark, 1990; Simonton, 2003; Thomke, 1998). As most engineering designs are constructed by modifying previous designs to meet new specifications, evaluators may resist designs that include too much novelty in terms of new components or new uses of existing components (Stacey & Eckert, 2010). Novel components have the potential to change the architecture of a design in ways that alter the interlinkages between the design's components, leading to more radical or architectural innovations (Henderson & Clark, 1990). In this vein, univalent expertise in the problem area may result in a preference for more feasible designs due to their greater familiarity with existing technologies compared to more novel, yet untested ones (Berlyne, 1960; Grant & Berry, 2011; Leonard-Barton, 1995; Mueller et al., 2012; Rindova & Petkova, 2007). However, given this understandable focus on feasibility, it is possible that the new possibilities generated by the novel components in a design—that can lead to changes to a design's architecture—will remain unnoticed, unrecognized, and underappreciated (Rindova & Petkova, 2007).

There are two potential yet countervailing effects of multivalent expertise on judgments of a design's feasibility, leading to two alternative hypotheses about multivalent experts' feasibility judgments relative to those of univalent experts. On one hand, the tension between a more novel yet unfamiliar design and a more feasible and reliable one may be strengthened as an evaluator's range of expertise increases in the problem area. We might expect that multivalent evaluators have broader and more sophisticated schemas and understandings of existing technologies in the relevant domains in the problem area. Accordingly, multivalent experts may be more likely to resist novel components, dismissing them as anomalies (Boudreau et al., 2016; Chai, 2017; Csikszentmihalyi, 1990). The uncertainty associated with novel components (Förster et al., 2009) may be even greater among multivalent than univalent evaluators if their expertise range enables them to see more weaknesses in the individual components used in a design (Stacey & Eckert, 2010). The primacy of engineering as a goal-oriented, problem-solving activity (Florman 1987, Vincenti 1990) may tilt a multivalent evaluator's preferences to be even more in favor of solutions that are perceived to be more feasible—and thereby reliable, stable, and implementable, but at the expense of lower novelty. Thus, we hypothesize:

Hypothesis H3a. Evaluators with multivalent expertise in the problem area are more likely than univalent evaluators to rate technological designs as higher in quality when they perceive the designs to be higher in feasibility and lower in novelty.

In contrast, multivalent expertise may be more willing to recognize the potential for novel components to improve a design's performance or functionality—as a system (Eckert et al., 2012). Since most engineering designs are constructed from existing technologies (Florman 1987), due to a dominant design for most established innovations (Henderson & Clark, 1990), novelty is likely to emerge when a designer cannot satisfy the needs (i.e., pre-specifications) of a task by modifying existing components or parts (Onarheim, 2012; Stacey & Eckert, 2010). As designers iterate and refine a design to satisfy the problem's pre-specifications, this process can lead to the development of novel parts that improve the functionality of the design as a workable device, thereby simultaneously increasing its feasibility. For example, in aerospace robotics, a simple device is often preferred over a complex one because it reduces the likelihood of malfunction or failure—which can have costly and tragic downsides (Chai et al. 2021). A novel approach that reduces the complexity of a proposed design, such as by combining components, eliminating subsidiary components, or the size of a component, can significantly improve the functionality of the system. Likewise, a lightweight design is preferable to a heavy one due to the performance enhancements, reduced cost, and increased payload mass capacity of lightweight structures. For instance, NASA's A-PUFFER robot, used to scout regions on the moon, is a lightweight, two-wheeled explorer robot. Its design initially came from Japanese origami, and the robot can flatten itself out and duck down to investigate tight spots that astronauts cannot get to on foot.[1]

Multivalent evaluators possessing architectural knowledge of the design's linkages and interconnections may more readily recognize the intuition, creativity, and ingenuity associated with a novel approach or component. Their multivalent expertise allows them to both detect a novel aspect in a design, and evaluate its functionality alongside other parts of the design (Henderson & Clark, 1990). Multivalent evaluators' architectural knowledge of the design, as a holistic system, may allow them to weigh the overall improvements in functionality offered by a novel design against the potential downsides and risks associated with untested technologies. In contrast, because univalent evaluators are limited in their domain knowledge of the problem area (only possessing expertise in a single domain), they will focus their attention on a design's components, where their expertise can be better applied (Camerer & Johnson, 1991). The insertion of a novel component may introduce uncertainty not only to the reliability of the component itself but also to its relationships with the other components in the design's architecture. Therefore, overall quality assessments will be highly correlated or consistent with a univalent evaluator's perceptions of a design's feasibility in its components or parts. Comparing the two forms of expertise we consider in the problem, we propose that multivalent expertise may allow evaluators to relax ex-ante assumptions about the architecture or form of the solution to consider alternative ways of solving a given design problem (Eckert et al., 2012). This greater appreciation of design novelty may attenuate multivalent evaluators' feasibility preference compared to univalent evaluators. Therefore, we offer an alternative hypothesis:

> *Hypothesis H3b. Evaluators with multivalent expertise in the problem area are less likely than univalent evaluators to rate technological designs as higher in quality when they perceive the designs to be higher in feasibility and lower in novelty.*

3. Research Design

In this section, we describe the setting and research design and provide details on the evaluator recruitment, procedures, random assignment of evaluators to solutions, 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 Astrobee 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

[1] https://spaceplace.nasa.gov/space-robots/en/

only humans can do. 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 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 evaluators from the community to help with solution evaluation for a subset of nine of the 17 challenges and all 101 purchased solutions within the nine Astrobee challenges. These challenges were selected due to their broad range of scope and complexity. 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) on 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 a 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 field of robotics that we pre-tested on individuals with different levels of expertise in the domain of robotics and related disciplines. The skills test included three categories of questions that drew upon the fields of mechanical engineering, electrical engineering, and computer science that collectively comprise the field of robotics engineering.

The focal problem area of the challenges is engineering design in aerospace robotics. Hence, aerospace and robotics engineering design correspond to the two types of relevant domain expertise in our setting. Table 1 provides a summary of the data collected and methods used to assess the registrants' domain expertise. As shown in Table 1, we have two types of univalent expertise in the problem area, namely domain expertise in robotics/mechatronics engineering and domain expertise in aerospace/defense. We assess evaluators' degree of robotics/mechatronics engineering knowledge using the robotics skills test and self-reported years of work experience in robotics/mechatronics. We assess the evaluators' aerospace/defense domain-relevant skills using their self-reported work experience (collected as a binary yes/no) in aerospace and defense. Given our interest in generating variation in evaluator expertise in the problem area, we recruited roughly equal numbers 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*, with a grade of 13 or more out of 17 (i.e., >75%), (iii) those who met the *human resources (HR) screen*, with at least two years of work experience in robotics/mechatronics engineering as well as work experience in aerospace/defense. 549 evaluators accepted our invitation to participate, and 374 completed the evaluation task. The skills test screen and HR screen represented two alternative approaches to assessing registrants' familiarity and experience in the problem area of engineering design in aerospace robotics. This produced 125 from the unscreened registrants, 140 from the skills test screen, and 109 from the HR screen.

Table 2 shows the distribution of evaluator attributes. The average score on the skills test is 10.37 (s.d. = 3.88) and the average number of years of work experience in robotics/mechatronics engineering is 1.66 (3.07). 13% of evaluators reported having some work experience in the aerospace/defense industry. Table 2 provides summary statistics on the evaluators. In terms of demographic characteristics, the evaluator pool is highly male, mostly between 25-34 years old on average, with the outside expertise evaluators being slightly younger, as well as 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.

As robotics is an interdisciplinary field, comprised of computer science, electrical engineering, and mechanical engineering, we note that the evaluators had a mean of 3.70 (s.d., = 6.75) years of work experience in software or computer science, a mean of 1.34 (s.d. = 3.32) years of work experience in electrical engineering, and a mean of 1.19 years (s.d. = 3.50) years of work experience in mechanical engineering, in a technical role. In addition, 65% of the evaluators had prior work experience within engineering or science organizations.

To validate our choices of domain expertise in robotics engineering design, we compare our expertise measures to the evaluator's self-perceived distance between their field of expertise and the field of robotics design using Jeppesen and Lakhani's (2010) measure of expertise distance. 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. The robotics univalent expertise evaluators had a mean rating of 1.87 (s.d. = 1.00), while the aerospace univalent expertise evaluators indicated a mean rating of 2.39 (s.d. = 0.980), which compares to a mean rating of 2.59 (s.d. = 1.36) across all evaluators. Importantly, we can draw a boundary on our evaluator pool as having at least some basic understanding and knowledge of robotics design problems.

[Tables 1 and 2 about here]

3.2. Evaluator Assignment and Evaluation Procedures

Our approach to assigning evaluators to solutions leverages judgments from multiple evaluators with a range of familiarity in the problem area, to correct for idiosyncrasies of any one evaluator, as well as having evaluators make comparative ratings of several solutions to achieve calibration along their continuum of scores (Amabile, 1982). Overall, our assignment of evaluators to solutions created 3,850 evaluator-solution pairs[2]. We used a randomized block design, in which each evaluator was first randomly assigned two of the nine challenges to evaluate. Then, within each challenge, we randomly assigned each individual five solutions 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 assigned to each evaluator while enabling them to compare across solutions within each challenge they were assigned to review. Overall, each solution received a mean of 38.3 (s.d. = 17.13) evaluations. In Table A1, we show that the randomized block design achieved balance across the evaluator covariates.

For each challenge, we gave evaluators a general overview of the challenge and then asked them to download and read the original problem statement (and associated pre-specifications) as well as familiarize themselves with the submission guidelines associated with the problem. Evaluators were told upfront that internal experts at NASA previously judged the solutions according to their feasibility and mass. After reading the challenge 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 scores and narrative comments corresponding to its feasibility, novelty, and overall quality, as well as report their confidence for each score. The narrative comments consisted of open-text responses where the evaluators were asked to document all the factors or aspects that led to their score (see Figure A1 for screenshots of evaluation procedures). Table 3 shows the mean number of evaluations per solution for each challenge.

[Table 3 about here]

3.3. Main Variables

3.3.1. Dependent Variables

To test H1 and H2, our main dependent variable is the *Feasibility score* assigned by reviewer i to solution j. We measured feasibility scores on a Likert scale from 1 (not at all feasible) to 7 (highly feasible). To

[2] There were 11 evaluators who completed more than one round of evaluations, which brings the total number of evaluations up from 3,740 (374 evaluators x 10 solutions) + (11 x 10) = 3,850 evaluator-solution pairs.

test H3 on the feasibility preference, the dependent variable is the *Quality score* assigned by reviewer *i* to solution *j*. Like feasibility, we measure quality on a Likert scale from 1 (lowest quality) to 7 (highest quality).

3.3.2. Explanatory Variables

Our main explanatory variable is the evaluator expertise type, either as univalent in robotics *or* aerospace, or multivalent in *both* robotics and aerospace. Consistent with our sampling strategy, we measure *Robotics univalent expertise* as any individual scoring a 13 or more out of 17 (>75%) on the robotics skills test *and* having two or more years of work experience in robotics/mechatronics within a technical organization. Figure A2 shows the distribution of skills test scores and years of robotics/mechatronics engineering work experience in our evaluator pool. We measure *Aerospace univalent expertise*, using a binary variable equal to 1 for all evaluators indicating that they have work experience in the domain of aerospace/defense and 0 otherwise. We did not collect the number of years of work experience in aerospace/defense in the demographic data collection.

To test H3 on the extent of the feasibility preference among univalent and multivalent expertise evaluators, we include two other explanatory variables, which are the *Feasibility score*, as described in section 3.3.1, and the *Novelty score* reviewer *i* assigns to solution *j*. Novelty is measured on a Likert scale from 1 (lowest novelty) to 7 (highest novelty) (see section 3.4 for details on the econometric approach to test H3).

3.3.3. 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 (101 in total). We use *Novelty confidence* and *Feasibility confidence* to control for the evaluator's confidence in their novelty and feasibility scores since prior work has noted that confidence varies with a decision-maker's level of expertise in the problem area (Kahneman and Klein 2009, Tversky and Kahneman 1974). We also control for several evaluator demographic characteristics (gender, age range, level of education, U.S. citizen/residing in the U.S.), which have been shown to affect decision-making (Weber and Johnson 2009).

3.4. Econometric Approach

We use ordinary least squares (OLS) models to estimate the relationship between solution feasibility and expertise range—univalent or multivalent in the problem area, as well as the relationship between feasibility preference and evaluator expertise range in the problem area. First, we investigate the relationship between the feasibility score for solution j and evaluator i's univalent expertise in (1):

$$Feasibility\ score_{ij} = \beta_0 + \beta_1 Robotics\ Univalent\ Expertise_i + \beta_2 Aerospace\ Univalent\ Expertise_i + \beta_3 Feasibility\ Confidence_{ij} + \beta_4 X_i + \gamma_j + \varepsilon_{ij}, \quad (1)$$

where X_i are evaluator covariates and γ_j are solution fixed effects. The solution fixed effects facilitate comparisons between evaluators randomly selected to evaluate the same solution, thereby facilitating within solution comparisons. We note that because expertise cannot be exogenously assigned, our estimation strategy relies on the random assignment of evaluators to solutions, and hence, exogenous variation in the intellectual overlap between evaluators and the solutions they are assigned to review. To investigate how feasibility scores are related to multivalent expertise in both robotics and aerospace, we add the interaction term between *High robotics expertise* x *High aerospace expertise* to (1).

Second, to investigate the relationship between feasibility preference and expertise range, we regress the *Quality score* for solution j on evaluator expertise, k (i.e., univalent expertise in robotics or aerospace/defense), *Feasibility score*, and *Novelty score*, as well as all unique two-way and three-way interactions between the different types of expertise and the component evaluation scores, as shown in (2):

$$Quality\ score_{ij} = \beta_0 + \beta_{1,k} \sum_k Univalent\ Expertise_{ik} + \beta_2 Feasibility\ score_{ij} + \beta_3 Novelty\ score_{ij} + \beta_4 \prod_k Univalent\ Expertise_{ik} + \beta_{5,k} \sum_k Univalent\ Expertise_{ik} \cdot Feasibility\ score_{ij} + \beta_{6,k} \sum_k Univalent\ Expertise_{ik} \cdot Novelty\ score_{ij} + \beta_7 \prod_k Univalent\ Expertise_{ik} \cdot Feasibility\ score_{ij} + \beta_8 \prod_k Univalent\ Expertise_{ik} \cdot Novelty\ score_{ij} + \beta_9 Feasibility\ confidence_{ij} + \beta_{10} Novelty\ confidence_{ij} + \beta_{11} X_i + \gamma_j + \varepsilon_{ij.} \quad (2)$$

Our primary coefficients of interest in (2) are β_7 and β_8, which correspond to the three-way interactions between the two types of univalent expertise and the feasibility and novelty scores, respectively.

4. Econometric Analyses: Experimental Results

Figure 2 plots the mean feasibility scores by the continuous measure of the robotics skills test score (Figure 2A), robotics years of work experience (Figure 2B), and aerospace work experience (Figure 2C), with standard errors. The red dashed line in Figures 2A and 2B illustrate the thresholds for the minimum scores on the skills test (i.e., 13 or above) and years of robotics work experience (i.e., 2 years or more) that we used to determine robotics univalent expertise in the problem area. The plots in Figure 2 provide consistent evidence that increasing univalent domain expertise leads to lower or more critical feasibility scores. In sections 4.1 and 4.2, we turn to test our main hypotheses.

[Figure 2 about here]

4.1. Perceptions of Feasibility and Expertise Range in the Problem Area (Hypotheses H1 and H2)

Hypothesis 1 theorized a negative relationship between feasibility scores and univalent domain expertise in the problem area, and H2 theorized that the slope of the relationship between feasibility scores and multivalent expertise would be less (H2a) or more (H2b) negative compared to univalent expertise. To test Hypothesis H1, in Table 4, Model 1, we begin by regressing *Feasibility scores* on *Robotics univalent* *expertise* and *Aerospace univalent expertise*, which are the two forms of univalent expertise in the problem area. We then add several controls in Models 2-4, including confidence ratings (Model 2), evaluator covariates (Model 3), and solution dummies (Model 4). Lastly, in Model 5, we test Hypothesis H2 by adding the interaction term between *Robotics univalent expertise* x *Aerospace univalent expertise*.

To test H1, Table 4, Model 1 shows that the coefficients for *Robotics univalent expertise* and *Aerospace univalent expertise* are both negative and significant (robotics: -0.415, $p < 0.01$; aerospace: -0.402, $p < 0.01$). These coefficients suggest that evaluators with univalent expertise in the problem area assign feasibility scores that are roughly 8.6 pp and 8.3 pp lower than outside evaluators without univalent expertise in the problem area. In Models 2-4, we show that these coefficients remain stable and robust to the confidence ratings (Model 2), evaluator covariates (Model 3), and solution dummies (Model 4).

To test H2, Table 4, Model 5 shows that the coefficient for the interaction term between *Robotics univalent expertise* x *Aerospace univalent expertise* is negative but not significant (Model 5: -0.0975*, ns).* Although the coefficient suggests that multivalent expertise evaluators are directionally more negative, the large standard errors indicate there is no strong evidence to indicate that multivalent expertise leads to comparatively lower feasibility scores relative to univalent expertise.

Taken together, we find support for H1, that univalent expertise leads to more negative (critical) feasibility scores, but we do not find support for H2—as there is no significant difference between the level of feasibility scores among univalent and multivalent expertise evaluators.

[Table 4 about here]

4.2. Perceptions of Quality and Expertise Range in the Problem Area (Hypothesis 3)

Hypothesis H3 theorized that the univalent expertise evaluators' preference for feasible solutions over novel ones would either be strengthened (H3a) or dampened (H3b) among multivalent expertise evaluators with expertise in both robotics and aerospace. In Table 5, Model 1, we begin by regressing *Quality scores* on both types of univalent expertise, as well as feasibility and novelty scores. Models 2 and 3 add the two-way interaction terms between the feasibility scores and *Robotics univalent expertise* (Model 2) and *Aerospace univalent expertise* (Model 3), and between the novelty scores and *Robotics univalent expertise* (Model 2) and *Aerospace univalent expertise* (Model 3). Model 4 adds both interaction terms from Models 2 and 3 into a single regression model. Next, to test H3, Model 5 adds the three-way interaction between *Robotics univalent expertise* x *Aerospace univalent expertise* x *Feasibility score* and *Robotics univalent expertise* x *Aerospace univalent expertise* x *Novelty score.* Lastly, Models 6-8 then add several covariates and controls for confidence ratings (Model 6), evaluator covariates (Model 7), and solution dummies (Model 8).

Turning to Model 1, we observe that both feasibility and novelty are positively associated with quality scores (feasibility score: 0.576, $p < 0.01$; novelty score: 0.362, $p < 0.01$), with the magnitude of the positive relationship with quality scores being roughly 1.6 times larger for feasibility than novelty.

Examining the interaction terms between the feasibility and novelty scores and *Robotics univalent expertise* in Model 2, we observe a positive and significant relationship between *Robotics univalent expertise* x *Feasibility score* (Model 2: 0.157, $p < 0.01$), and a negative and significant relationship between *Robotics univalent expertise* x *Novelty score* (Model 2: -0.0872, $p < 0.05$). These coefficients suggest that relative to outside evaluators, evaluators with robotics univalent expertise weight feasibility 28.3% more (0.157/0.553 x 100) and novelty 23.3% less (-0.0872/0.374 x 100) in their assessments of a design's overall quality. Similarly, the interaction terms between the feasibility and novelty scores and *Aerospace univalent expertise* in Model 3 indicate a positive and significant relationship between *Aerospace univalent expertise* x *Feasibility score* (Model 3: 0.159, $p < 0.01$) and a negative and significant relationship between *Aerospace univalent expertise* x *Novelty score* (Model 3: -0.115, $p < 0.01$). These coefficients can be interpreted as follows: aerospace univalent expertise evaluators assign 28.7% more weight (0.159/0.554 x 100) to feasibility and 30.5% less weight (-0.115/0.377 x 100) to novelty in their evaluations of a design's quality. Taken together, the coefficients of the interaction terms in both Models 2 and 3 suggest that univalent expertise in either robotics (Model 2) or aerospace (Model 3) causes evaluators to overweight feasibility relative to novelty when making overall judgments of a solution's quality.

Next, to examine how multivalent expertise affects the relationships between the observed feasibility preference among univalent expertise evaluators, in Model 5 we examine the three-way interaction between *Robotics univalent expertise* x *Aerospace univalent expertise* x *Feasibility score* and *Robotics univalent expertise* x *Aerospace univalent expertise* x *Novelty score.* We find patterns consistent with attenuation of the feasibility preference (i.e., H3b): the three-way interaction between multivalent expertise and feasibility scores is negative and significant (Model 5: -0.192, $p < 0.05$) and directionally positive between multivalent expertise and novelty scores (Model 5: 0.134, *ns*). Since evaluators were randomly assigned to solutions, the coefficients remain consistent in Models 6-8, with the addition of various controls for confidence (Model 6), evaluator covariates (Model 7), and solution fixed effects (Model 8).

[Table 5 about here]

Overall, we find evidence in support of Hypothesis H3b: multivalent expertise attenuates the feasibility preference among univalent expertise evaluators for solutions that are higher in feasibility but lower in novelty.

5. Text Analysis: Topic Modeling and Sentiment Analysis of Evaluator Comments

Immediately following their ratings of the feasibility, novelty, and quality, we asked each evaluator to document the factors contributing to their assigned scores. To examine potential differences in the evaluators' comments by expertise type, we perform two complementary types of text analysis.

First, we apply latent Dirichlet allocation (LDA), an algorithmic method for uncovering latent topics in a corpus of data (Blei et al. 2003), on the open-text feasibility comments. We chose to focus on the feasibility comments due to the relative importance of feasibility in engineering design (Dean et al., 2006; Shah et al., 2003; Vincenti, 1990), and our goal of understanding the extent that univalent and multivalent expertise affects evaluator assessments of the feasibility of a focal design.[3] Topic modeling uses the co-location of words in a collection of documents to learn about the unobserved topic structure and compute the distribution of topics per document, as well as the distribution of words over topics (Blei et al. 2003). While the documents and the words in the documents are observed, the topics represent the "hidden structure" and are not observed (Blei 2012). Computationally, the algorithm identifies the posterior distribution of observed variables in a collection of documents. One key benefit of LDA is that it does not require classification by humans, so the topics are not influenced by the semantic matters that researchers expect to find, instead having the structure emerge from the data. The researcher can infer the meaning of each of these topics and represent each document as a probabilistic distribution over the possible topics. We estimate the model on our feasibility and novelty comments from the sample of evaluators using the *topicmodels* package in R (Hornik and Grün 2011).[4] One crucial choice when using LDA is the number of topics to be estimated by the algorithm. We selected the number of topics by examining the model fit over a range of topic numbers, from two to nine.

Second, we perform sentiment analysis on the evaluators' comments to identify differences in sentiment towards their feasibility, novelty, or quality scores by expertise type in the problem area. Sentiment analysis is the use of natural language processing (NLP) to systematically identify positive or negative valence from textual data. We use TextBlob (Loria, 2018) to perform sentiment analysis of the interns' preprocessed responses. TextBlob refers to a lexicon called en-sentiment.xml and returns a polarity score from -1 to 1 for each word in its lexicon, where -1 defines a negative sentiment and 1 defines a positive sentiment. The polarity of a longer text is then computed as the average polarity of all words and phrases in the data.

5.1. Topic Modeling Results of Feasibility Comments

We fit a LDA model using three topics on the corpus of feasibility comments (N = 3,445 comments), which returned the best goodness-of-fit in terms of the perplexity of a held-out set of documents. Topic 1 includes words corresponding to *Robotics-specific criteria*, with sample words including "arm", "architecture", "position", "power", "robotic", "movement", and "mechanics". Topic 2 includes words pertinent to the *Aerospace-specific criteria*, such as "mass", "force", "material", "rubber" and "friction". Topic 3 includes

[3] We mention that we also performed LDA on the novelty and quality comments but did not detect semantic differences between topics.

[4] We examined the distribution of words across topics for the novelty and quality comments but did not find any semantic differences between topics.

words associated with the *Systems-level criteria*, such as “detailed”, “simple”, “easy”, “understand”, “components”, and “system”. Table 6 provides sample comments by topic classified using the LDA model. Whereas the first two topics, namely the robotics- or aerospace-specific criteria tend to focus on the details of the design components, the third topic, *Systems-level criteria*, tends to offer more holistic assessments of the design’s overall functionality, such as emphasizing the interconnections between the design’s components as opposed to commenting on the specific components or parts.

[Table 6 about here]

We examine the distribution of topics across the evaluators’ feasibility comments by using OLS models to regress the proportion of topic t found in each comment on evaluator i’s expertise type for solution j:

$$Topic\ Proportion_{\ ijt} = \beta_0 + \beta_1 High\ Robotics\ Expertise_i + \beta_2 High\ Aerospace\ expertise_i + \beta_3 High\ Robotics\ Expertise_i \times High\ Aerospace\ expertise_i + \varepsilon_{ijt} \text{ for } t = 1,2,3. \quad (4)$$

Table 7 reports the OLS results of the proportional representation of *Robotics-specific criteria* (Topic 1; Models 1-2), *Aerospace-specific criteria* (Topic 2; Models 3-4), and *Systems-level criteria* (Topic 3; Models 5-6) on evaluator expertise type. In Model 1, we observe that evaluators with univalent expertise in robotics are 6.24 pp more likely to comment on *Robotics-specific criteria* compared to outside expertise evaluators (Model 1: 0.0624, p < 0.01), and there is no difference between the aerospace univalent and outside evaluators (Model 1: 0.00848, *ns*). The interaction term between the two types of univalent expertise in Model 2 also indicates that multivalent expertise does not have a differentiated effect on the likelihood of commenting on the *Robotics-specific criteria* (Model 2: -0.0222, *ns*).

Turning to the *Aerospace-specific criteria* (Topic 2; Models 3-4), while there is no significant difference in Model 3 among the univalent evaluators in the likelihood of commenting on aerospace-specific criteria (robotics: 0.00745, *ns*; aerospace: -0.00952, *ns*), the interaction term between *Robotics univalent expertise* x *Aerospace univalent expertise* in Model 4 is negative and highly significant (Model 4: -0.145, p < 0.01), suggesting that the multivalent evaluators are less likely than univalent expertise evaluators to comment on the specific aerospace-relevant components of the design. Lastly, turning to the proportion of *Systems-level criteria* (Topic 3) in each comment, we find that while the robotics univalent evaluators are significantly less likely to provide systems-level comments about the design compared to outside expertise evaluators (Model 5: -0.0698, p < 0.01) and there is no difference between the aerospace univalent and outside evaluators (Model 5: 0.00104, *ns*), the interaction term between the two types of univalent expertise is positive and significant (Model 6: 0.167, p < 0.01). The magnitude of the coefficient suggests almost half (i.e., 48.3%) of the multivalent evaluators’ comments focus on systems-level criteria, which is 41.1% and 14.3% more than the robotics and aerospace univalent expertise evaluators, respectively. We note that the regression results reported in Table 7 remain robust to solution fixed effects (see Table A3).

Overall, the topic modeling results of the feasibility comments reveal that expertise range in the problem area affects the extent that evaluators perceive a design by its components or as a system. The LDA topic modeling results suggest that whereas univalent evaluators are more likely to focus their assessments on the specific components of the design, in line with their domain of expertise in the problem area, multivalent expertise evaluators are more likely to judge the design's architecture as a system, indicating a shift in attention from detailed- to holistic-oriented. Although the differences in attention allocation are larger between the robotics univalent and multivalent evaluators compared to the aerospace univalent and multivalent evaluators, one plausible explanation can be attributed to the fact that our evaluation pool is highly educated in technically-oriented fields, particularly in computer science/software engineering, mechanical engineering, and electrical engineering—which are the three engineering disciplines that comprise robotics/mechatronics engineering (see section 3.1). This suggests that the aerospace univalent expertise evaluators were likely to have at least some familiarity with the robotics-specific criteria of the designs. This is substantiated by the relatively higher skills test score among the aerospace univalent expertise evaluators (mean = 12.13 or 71.35%, s.d. = 2.93), compared to a mean score of 10.05 or 59.12% (s.d. = 3.95) for evaluators without aerospace univalent expertise ($t = 11.638$ $p < 0.001$). That being said, the evaluator pool, on the whole, did comparatively better than pure chance on the skills test (i.e., 3.4 or 20%, as the questions were multiple choice, each with 5 possible responses).

5.2. Sentiment Analysis Results

Next, we examine the relationships between the polarity of the evaluators' feasibility, novelty, and quality comments and expertise type or range in the problem area using OLS models to regress comment polarity on evaluator i's expertise type for solution j:

$$Comment\ polarity_{ij} = \beta_0 + \beta_1 Robotics\ Univalent\ Expertise_i + \beta_2 Aerospace\ Univalent\ Expertise_i + \beta_3 Robotics\ Univalent\ Expertise_i \times Aerospace\ Univalent\ Expertise_i + \varepsilon_{ij}. \quad (3)$$

Table 8 reports the results from the OLS models estimating the relationships between the polarity of feasibility (Models 1-2), novelty (Models 3-4), and quality comments (Models 5-6) on evaluator expertise type—i.e., univalent, or multivalent expertise in the problem area. Model 1 shows that robotics univalent evaluators' feasibility comments are more critical compared to outside expertise evaluators (Model 1: -0.0280, $p < 0.05$), and that there is no difference in feasibility comment polarity among the aerospace univalent and outside evaluators (Model 1: -0.00436, *ns*). In Model 2, the interaction term between the two types of univalent expertise is also not significant (Model 2: 0.0169, *ns*). Overall, the polarity of the evaluators' feasibility comments is consistent with the econometric results presented in Table 4, which showed that having some knowledge overlap in the domain leads to lower scores and less positive comments, but there is no marginal effect of multivalent expertise on either the level of scores or valence of the comments.

Turning to the novelty comments in Model 3, we find that the relationship between the polarity of the novelty comments and univalent expertise is directionally negative (robotics univalent: -0.0194, *ns*; aerospace univalent: -0.0190, $p < 0.10$). This said Model 4 shows a positive and significant interaction effect between *Robotics univalent expertise* x *Aerospace univalent expertise* (Model 4: 0.0572, $p < 0.05$). The significant increase in positive sentiment compared to the univalent evaluators is consistent with the notion that multivalent evaluators may be more likely to recognize the potential performance improvements enabled by novel components or features. Lastly, examining the quality comments, whereas Model 5 shows a negative and significant effect of both kinds of univalent expertise on the polarity of the quality comments (robotics univalent: -0.0470, $p < 0.01$; aerospace univalent: -0.0364, $p < 0.01$), Model 6 indicates that multivalent evaluators are significantly more positively valenced about a design's overall quality (Model 6: 0.115, $p < 0.01$). The observed patterns relating the sentiment of the quality comments and expertise type in the problem area suggest that the greater range of knowledge among the multivalent evaluators is associated with a lesser focus on the specific details of a design's components, and potentially more holistic assessments of a design's architecture and functionality. In Tables A2 and A3, we show that the reported results in Tables 7 and 8 are robust to solution fixed effects.

[Table 8 about here]

6. Does Evaluator Expertise Lead to Substantively Different Decisions?

In this section, we examine the extent that the evaluators' differences in perceptions by expertise type (i.e., range) result in substantive differences in the rank-ordering of the solutions. On one hand, it is possible that expertise type alters what evaluators perceive about a design's characteristics but does not change their overall assessments of a design's quality. On the other hand, expertise type may alter both evaluator perceptions and assessments of a design's quality. To address this question, we first rank-order the solutions by their mean score within each challenge, among the different types of expertise in the problem area. Figure 3 presents three scatterplots of the within challenge-solution quality rankings between the aerospace univalent evaluators (y-axis) and robotics univalent expertise evaluators (x-axis) (A), followed by the quality rankings between the multivalent expertise (y-axis) and robotics univalent expertise evaluators (x-axis) (B), and the multivalent expertise (y-axis) and aerospace univalent expertise evaluators (x-axis) (C). The red dashed line is the 45-degree line, which corresponds to rank orders where the two expertise types agreed on the solution rank orders. The blue line corresponds to the non-parametric LOESS regression.

In Figure 3A, we observe that very few points fall on the red dashed 45-degree line, suggesting that there is considerable disagreement in the rank-ordering of solution quality within each challenge by the two types of univalent expertise in the problem area. Indeed, the Spearman rank-order correlation between the two types of univalent expertise is only moderately correlated at $\rho = 0.58$. The mean absolute difference in rankings between robotics and aerospace univalent evaluators is 2.83 (s.d. = 2.93). Examining the distribution of these differences in rank orders by expertise type, we observe that whereas 16 (15.8%) of the designs are unchanged between the aerospace and robotics univalent evaluators, 47 (46.5%) are within three rank orders of each other, and 38 (37.6%) are three or more rank-orders apart. The sizable turnover in rank orders between the robotics and aerospace univalent evaluators suggests that each type of univalent expertise may be perceiving different information cues about the components of the design, depending on their specific domain of knowledge overlap with the problem area.

[Figure 3 about here]

In comparison to Figure 3A, the Spearman rank-order correlations in Figure 3B and 2C of $\rho = 0.65$ and $\rho = 0.88$ indicate that there is a greater agreement between the univalent and multivalent expertise evaluators than between the different types of univalent evaluators. This may be expected given that the multivalent evaluators have both forms of univalent expertise in their knowledge base. This said, the higher Spearman rank-order correlations between the aerospace univalent evaluators and the multivalent evaluators (Figure 3B) compared to the robotics univalent and multivalent evaluators (Figure 3C) once again reflects the highly technical nature of our evaluator pool.

Overall, the Spearman correlations in Figure 3 illustrate that evaluators' expertise type in the problem area, as either univalent or multivalent, can lead to substantive differences in proposal rankings and subsequent resource allocation decisions, due to differences in how prior knowledge of the problem area shapes an evaluator's perceptions of a design, either as its components or as a system.

7. Discussion and Conclusion

A fundamental challenge in the evaluation and selection of early-stage technological innovations is how to effectively draw upon individuals with the relevant domain expertise to judge the overall quality of novel designs and solutions that can be feasibly implemented into a workable product or device. Often, technological innovations draw upon knowledge from multiple domains, meaning that different kinds of expertise may be required to evaluate the overall quality of alternative designs. Yet we know relatively little about how the range of an evaluator's expertise in the problem area—i.e., univalent in a single domain of the problem area or multivalent, spanning multiple domains in the problem area—is systematically related to evaluators' perceptions of quality. Drawing on the notion that technological innovations require both component knowledge of the parts in a design and architectural knowledge of how the components or parts are integrated into a holistic design (Clark, 1985; Henderson & Clark, 1990; Vincenti, 1990), we propose that an evaluator's expertise range in the problem area relates to their perceptions of the design either by its parts or as a whole, to shape their evaluations of overall quality. To examine these relationships, we collaborated with NASA and Freelancer.com to design a field experiment, recruiting a diverse range of evaluators with different types of expertise in the problem area of aerospace robotics to evaluate engineering designs solicited from an innovation challenge series. We collected intricate background information on the evaluators, including demographic data and a robotics skills test assessment, which were used to identify two types of domain-relevant knowledge in aerospace robotics design, namely univalent expertise in either aerospace or robotics or multivalent expertise in aerospace and robotics. We randomly assigned evaluators to solutions, creating exogenous variation between the evaluators' range of expertise and the solutions they were assigned to evaluate. This approach facilitated causal estimates of how expertise range relates to evaluation scores while holding the characteristics of the evaluators and solutions constant.

We report several noteworthy patterns. First, our experimental results show that, compared to non-expert ("outside") evaluators, evaluators with either univalent or multivalent expertise provide harsher (i.e., more critical) judgments of design feasibility; univalent and multivalent experts do not differ in the comparisons to non-experts. Second, we observe that univalent evaluators exhibit a *feasibility preference*, in which evaluators with knowledge overlap in a single domain of the problem area (either robotics or aerospace, but not both), are more likely to perceive solutions as higher in overall quality when they are higher in their perceived feasibility and lower in novelty.

Third, and most notably, our main finding indicates that the feasibility preference is attenuated in evaluators possessing multivalent expertise spanning both domains of the problem area. That is, multivalent evaluators are more likely than univalent evaluators to perceive that a high-quality design is a novel solution that can be feasibly implemented into a workable device. Lastly, we perform natural language processing of the evaluators' open-text comments to gain deeper insights into the observed patterns. Topic modeling and sentiment analyses of the evaluators' comments indicate that expertise range alters how evaluators perceive a design's features. Whereas univalent expertise enables evaluators to see more *details* about a design's components, allowing them to uncover more weaknesses in a design's feasibility relative to outside evaluators, multivalent expertise facilitates more *holistic* assessments of a design's architecture—i.e., the interconnections between its components, facilitating assessments of the design's functionality. This multivalent expertise enables these evaluators to recognize and appreciate the potential upsides associated with more novel and feasible designs.

Our study contributes primarily to our understanding of how technological innovations emerge, by drawing a connection between evaluator expertise and preferences for different kinds of innovation. In particular, our research builds on recent work that investigates the effect of expertise in the problem area on evaluation outcomes (Boudreau et al., 2016; Criscuolo et al., 2017; Lane et al., 2021; Li, 2017). Whereas prior work focuses on the intellectual distance between an evaluator's expertise and the knowledge embodied in the proposal or idea as a single dimension, there is a potential gap in our understanding of these relationships, as technological innovations often require knowledge inputs from multiple domains.

To this end, we conceptualize an evaluator's expertise in the technological problem area as being multi-dimensional in nature, differentiating between univalent and multivalent expertise (i.e., domain-relevant knowledge of a single domain versus multiple domains), as well as their relationships with evaluators' perceptions of overall quality. Our findings suggest that an evaluator's range of expertise in the problem area is associated with whether an evaluator perceives a design as a collection of components, or as a holistic system. These findings have implications for the selection of experts to evaluate early-stage technological ideas. Whereas different forms of univalent expertise may lead to more discerning and comprehensive judgments of a design's feasibility, a potential downside is that it may result in overly conservative decisions favoring incremental improvements that prioritize feasibility over novelty (Lane et al., 2021). In contrast, multivalent expertise favors more balanced designs (in terms of their novelty and feasibility), that have greater potential to offer more substantial improvements to a design's functionality. While a natural implication of our findings for evaluations of technological designs is to recruit both univalent and multivalent evaluators to provide complementary perspectives on a design, it is also important to recognize that a firm's technological trajectory is a strategic decision, meaning that the composition of evaluation panels or independent evaluators (Lane et al., 2021) ought to be informed by the needs of the problem or task at hand. Our work suggests that univalent and multivalent evaluators' alternative perspectives of a design can result in different resource allocation decisions, depending on whose perspectives are considered. If the design task is relatively routine (i.e., continues along the existing technological trajectory), then univalent evaluators may be better suited for evaluating the extent to which a focal design's components are feasible given a pre-specified set of constraints. Alternatively, if the design task calls for more novel and untested solutions, then multivalent evaluators may be a better match to evaluate and select designs that depart more substantially from the existing technological frontier and challenge the existing architecture of the dominant design (Abernathy & Utterback, 1978; Albert & Siggelkow, 2022; Henderson & Clark, 1990).

Our work opens the door for future work on the evaluation and selection processes of technological ideas, particularly on how opinions across multiple individuals with diverse perspectives can be combined for decision-making purposes. Our study suggests that the community is relatively efficient in self-sorting in response to a broadcast call for evaluators in a technical skills-intensive technological problem area. This is exemplified by the considerably high scores on the robotics skills test across the entire evaluator pool (mean = 10.34 out of 17, s.d. = 3.90), and the high percentage of participants—i.e., 87 percent—having some previous work experience in a technical organization (e.g., electrical, mechanical, manufacturing, robotics/mechatronics engineering, etc.). The efficiency with which evaluators self-sorted into our evaluator pool offers new possibilities for investigating how external evaluators and evaluators from analogous domains (Dahlander et al. 2016, Franke et al. 2014, Gieryn 1983, Jeppesen and Lakhani 2010) might introduce valuable heterogeneity in perspectives to evaluation processes that traditionally rest in the hands of a few internal experts (Boudreau et al., 2016). Given that experts' mental maps are fragile, and perform poorly outside of their prior experiences (Budescu & Chen, 2015; Camerer & Johnson, 1991), the use of "crowd" or community evaluators, in aggregate, can offer alternative yet useful perspectives to be considered by internal experts. This approach may be advantageous in recruiting additional multivalent evaluators, whose expertise range spans the domains of the problem area, and tends to be rare among internal experts—due to their embeddedness in the existing organizational routines and structures of the firm (Henderson & Clark, 1990).

Moreover, although we focused primarily on how expertise range shapes evaluations of technological ideas, a critical next phase is the role of the aggregator and the use of alternative decision rules to make resource allocation decisions (Christensen & Knudsen, 2010; Csaszar & Eggers, 2013). Indeed, past studies have documented the significant "noise" and low interrater reliability among expert evaluators (Cole and Simon 1981, Pier et al. 2018, Rothwell and Martyn 2000). Given that expertise may shape how and what evaluators perceive and prioritize during their evaluations, simple averaging efforts across evaluators with diverse expertise types may not be conducive to decision-making due to the differences in how evaluators are allocating their attention. The opening up of the evaluation process to individuals with heterogeneous expertise and backgrounds, from outside the firm's boundaries (Franzoni & Sauermann, 2014; Greenberg & Mollick, 2017; Mollick & Nanda, 2016), creates new opportunities to consider alternative evaluation structures (e.g., hierarchy, polyarchy, hybrid) (Christensen & Knudsen, 2010) and their implications on aggregation decisions (Csaszar & Eggers, 2013).

Lastly, in our work, we draw a connection between the problem formulation stage (Baer et al., 2013; Von Hippel & Von Krogh, 2016) and evaluations of alternative solutions aimed at addressing the problem at hand. As problem pre-specifications draw a boundary around the range of solutions that might be considered in the design space, the number and type of pre-specifications on the initial problem may not only affect the types of solutions generated (Acar et al., 2019) but also the perceived quality of alternative solutions during evaluations of them. Whereas the design literature has established that creativity can emerge as a result of the number of pre-specifications imposed on the problem (Onarheim, 2012; Stacey & Eckert, 2010), one potential challenge that we raise with this approach is that it may create a fixation with feasibility in the evaluation stage. Univalent evaluators with a componential view of the problem area may overlook novel designs due to the importance of identifying a feasible solution. Perhaps an alternative approach might be to reduce the number of pre-specifications or incorporate novelty as a desired pre-specification at the problem formulation stage. For instance, many design problems include targets for the percentage of components to be reused from existing technologies or products (Eckert et al., 2012). Rephrasing these targets in terms of the percentage of novel components that differ from those in existence may prompt evaluators to perceive novelty more favorably as a feature to be desired rather than avoided. These subtle changes to the problem formulation may encourage evaluators to recognize and appreciate novelty in proposed solutions, which is likely to have direct implications on the types of innovations that are then selected for implementation. A novelty pre-specification, might, for instance, lead to the identification of novel yet feasible innovations that introduce fundamental changes to the architecture of existing technologies (Dewar & Dutton, 1986; Ettlie et al., 1984; Henderson & Clark, 1990).

Overall, our study enrichens understanding of how expertise range in the problem area drives support for innovations that continue to expand the technological frontier in novel and feasible directions.

8. References

Acar OA, Tarakci M, van Knippenberg D (2019) Creativity and innovation under constraints: A cross-disciplinary integrative review. *J. Manag.* 45(1):96–121.

Amabile TM (1983) The social psychology of creativity: A componential conceptualization. *J. Pers. Soc. Psychol.* 45(2):357.

Anders KE, Charness N, Feltovich P, Hoffman RR (2006) *Cambridge handbook of expertise and expert performance* (Cambridge University Press).

Askin N, Mauskapf M (2017) What makes popular culture popular? Product features and optimal differentiation in music. *Am. Sociol. Rev.* 82(5):910–944.

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, Fons-Rosen C, Graff Zivin JS (2019) Does science advance one funeral at a time? *Am. Econ. Rev.* 109(8):2889–2920.

Baer M, Dirks KT, Nickerson JA (2013) Microfoundations of strategic problem formulation. *Strateg. Manag. J.* 34(2):197–214.

Barnett SM, Koslowski B (2002) Adaptive expertise: Effects of type of experience and the level of theoretical understanding it generates. *Think. Reason.* 8(4):237–267.

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

Berlyne DE (1960) Conflict, arousal, and curiosity.

Blei DM (2012) Probabilistic topic models. *Commun. ACM* 55(4):77–84.

Blei DM, Ng AY, Jordan MI (2003) Latent dirichlet allocation. *J. Mach. Learn. Res.* 3(Jan):993–1022.

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.

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.

Carson RT, Graff Zivin J, Louviere JJ, Sadoff S, Shrader JG (2022) The Risk of Caution: Evidence from an Experiment. *Manag. Sci.*

Catalini C (2018) Microgeography and the direction of inventive activity. *Manag. Sci.* 64(9):4348–4364.

Cattani G, Deichmann D, Ferriani S (2022) Novelty: Searching for, seeing, and sustaining it. *Gener. Recognit. Legitimation Nov.* (Emerald Publishing Limited).

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.*

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

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.

Csikszentmihalyi M (1990) The domain of creativity.
Dahlander L, Jeppesen LB, Piezunka H (2019) How organizations manage crowds: Define, broadcast, attract, and select. *Manag. Inter-Organ. Collab. Process Views Res. Sociol. Organ. Vol 64 Emerald Publ. Ltd.*:239–270.
Dane E (2010) Reconsidering the trade-off between expertise and flexibility: A cognitive entrenchment perspective. *Acad. Manage. Rev.* 35(4):579–603.
Dean DL, Hender J, Rodgers T, Santanen E (2006) Identifying good ideas: constructs and scales for idea evaluation. *J. Assoc. Inf. Syst.* 7(10):646–699.
Dewar RD, Dutton JE (1986) The adoption of radical and incremental innovations: An empirical analysis. *Manag. Sci.* 32(11):1422–1433.
Diehl M, Stroebe W (1987) Productivity loss in brainstorming groups: Toward the solution of a riddle. *J. Pers. Soc. Psychol.* 53(3):497.
Falchetti D, Cattani G, Ferriani S (2022) Start with "Why," but only if you have to: The strategic framing of novel ideas across different audiences. *Strateg. Manag. J.* 43(1):130–159.
Ferioli M, Dekoninck E, Culley S, Roussel B, Renaud J (2010) Understanding the rapid evaluation of innovative ideas in the early stages of design. *Int. J. Prod. Dev.* 12(1):67–83.
Florman SC (1987) *The civilized engineer* (St. Martin's Griffin).
Förster J, Liberman N, Shapira O (2009) Preparing for novel versus familiar events: Shifts in global and local processing. *J. Exp. Psychol. Gen.* 138(3):383.
Franzoni C, Sauermann H (2014) Crowd science: The organization of scientific research in open collaborative projects. *Res. Policy* 43(1):1–20.
Goldberg A, Hannan MT, Kovács B (2016) What does it mean to span cultural boundaries? Variety and atypicality in cultural consumption. *Am. Sociol. Rev.* 81(2):215–241.
Grant AM, Berry JW (2011) The necessity of others is the mother of invention: Intrinsic and prosocial motivations, perspective taking, and creativity. *Acad. Manage. J.* 54(1):73–96.
Greenberg J, Mollick E (2017) Activist choice homophily and the crowdfunding of female founders. *Adm. Sci. Q.* 62(2):341–374.
Harvey S, Mueller JS (2021) Staying alive: Toward a diverging consensus model of overcoming a bias against novelty in groups. *Organ. Sci.* 32(2):293–314.
Hasan S, Koning R (2019) Conversations and idea generation: Evidence from a field experiment. *Res. Policy* 48(9):103811.
Henderson R, Clark KB (1990) Architectural innovation: The reconfiguration of existing product technologies and the failure of established firms. *Adm. Sci. Q.*:9–30.
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).
Hornik K, Grün B (2011) topicmodels: An R package for fitting topic models. *J. Stat. Softw.* 40(13):1–30.
Jeppesen LB, Lakhani KR (2010) Marginality and problem-solving effectiveness in broadcast search. *Organ. Sci.* 21(5):1016–1033.
Jones BF (2009) The burden of knowledge and the "death of the renaissance man": Is innovation getting harder? *Rev. Econ. Stud.* 76(1):283–317.
Kahneman D, Klein G (2009) Conditions for intuitive expertise: a failure to disagree. *Am. Psychol.* 64(6):515.
Kahneman D, Tversky A (1979) Prospect theory: An analysis of decision under risk. *Econometrica* 47(2):363–391.
Koen BV (1985) *Definition of the Engineering Method.* (ERIC).
Kuhn TS (1962) 9. The Essential Tension: Tradition and Innovation in Scientific Research. *Essent. Tens.* (University of Chicago Press), 225–239.
Lamont M (2009) *How professors think* (Harvard University Press).

Lane J, Ganguli I, Gaule P, Guinan E, Lakhani KR (2021) Engineering serendipity: When does knowledge sharing lead to knowledge production? *Strateg. Manag. J.* 42(6).

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.*

Langfeldt L (2004) Expert panels evaluating research: decision-making and sources of bias. *Res. Eval.* 13(1):51–62.

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

Lee CJ (2015) Commensuration bias in peer review. *Philos. Sci.* 82(5):1272–1283.

Levinthal DA (1997) Adaptation on rugged landscapes. *Manag. Sci.* 43(7):934–950.

Lobo A (2020) Case Study of Toyota Prius- The world's most popular Hybrid Electric Vehicle.

Malone-Kircher M (2016) James Dyson on 5,126 Vacuums That Didn't Work–and the One That Finally Did. *N. Y.* 22.

Medin DL, Lynch EB, Coley JD, Atran S (1997) Categorization and reasoning among tree experts: Do all roads lead to Rome? *Cognit. Psychol.* 32(1):49–96.

Miron-Spektor E, Erez M (2017) *Looking at creativity through a paradox lens* (Oxford University Press Oxford).

Mollick E (2014) The dynamics of crowdfunding: An exploratory study. *J. Bus. Ventur.* 29(1):1–16.

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

Mount MP, Baer M, Lupoli MJ (2021) Quantum leaps or baby steps? Expertise distance, construal level, and the propensity to invest in novel technological ideas. *Strateg. Manag. J.* 42(8):1490–1515.

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

Myers K (2020) The elasticity of science. *Am. Econ. J. Appl. Econ.* 12(4):103–34.

Newell A (1979) Reasoning, problem solving and decision processes: The problem space as a fundamental category.

Newell A, Simon HA (1972) *Human problem solving* (Prentice-Hall Englewood Cliffs, NJ).

Noor A (2012) Emerging interdisciplinary fields in the coming intelligence/convergence era. *Open Eng.* 2(3):315–324.

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

Shah JJ, Smith SM, Vargas-Hernandez N (2003) Metrics for measuring ideation effectiveness. *Des. Stud.* 24(2):111–134.

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

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

Simonton DK (1999) *Origins of genius: Darwinian perspectives on creativity* (Oxford University Press).

Simonton DK (2003) Scientific creativity as constrained stochastic behavior: the integration of product, person, and process perspectives. *Psychol. Bull.* 129(4):475.

Stokes DE (1997) *Pasteur's quadrant: Basic science and technological innovation* (Brookings Institution Press).

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

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

Von Hippel E, Von Krogh G (2016) Crossroads—Identifying viable "need–solution pairs": Problem solving without problem formulation. *Organ. Sci.* 27(1):207–221.

Wagner RK (1987) Tacit knowledge in everyday intelligent behavior. *J. Pers. Soc. Psychol.* 52(6):1236.

Wagner RK, Sternberg RJ (1985) Practical intelligence in real-world pursuits: The role of tacit knowledge. *J. Pers. Soc. Psychol.* 49(2):436.
Whitley R (2000) *The intellectual and social organization of the sciences* (Oxford University Press on Demand).

Table 1. Methods for Assessing Evaluator Domain Expertise in the Problem Area

Evaluator domain expertise	Data Collection Method
Robotics/mechatronics engineering	• 17-question skills test in robotics assessing knowledge in mechanical engineering, electrical engineering, and computer science equivalent to undergraduate level courses in robotics • Self-reported years of work experience in robotics/mechatronics engineering (outside of educational experience)
Aerospace/defense	• Self-reported work experience in the aerospace/defense industry

Table 2. Summary Statistics on Evaluator Demographics (N = 374)

Variable	Mean (s.d.)	Min	Max
Skills test score (out of 17)	10.37 (3.88)	0	17
Robotics years of experience	1.66 (3.07)	0	20
Aerospace/Defense	0.13	0	1
Female	0.14	0	1
Age range	25 - 34 (Mode)	Under 18	65 or older
Bachelors	0.76	0	1
Masters	0.25	0	1
USA	0.067 (0.251)	0	1
Robotics design expertise	2.59 (1.36)	1	5

Note: age categories are under 18 , 18-24, 25-34, 35-44, 45-54, 55-64, 65 or older, prefer not to say

Table 3. Mean (s.d.) of Evaluator-Solution Pairs by Challenge and Evaluator Expertise Type (N = 3,850)

Challenge	# of Solutions	Overall
HMSA	6	72.7 (1.63)
MDC	14	28.2 (2.39)
MIS	11	41 (2.45)
PSA	6	71.7 (2.16)
RASA	6	68.5 (2.43)
SAM	12	41.2 (3.11)
SDM	18	21.3 (2.54)
SPAM	12	37.8 (2.04)
SRA	16	25.9 (4.70)
Total	101	38.3 (17.13)

Table 4. OLS Models of Feasibility Scores on Evaluator Univalent and Multivalent Expertise

	Dependent Variable: Feasibility Rating				
VARIABLES	Model 1	Model 2	Model 3	Model 4	Model 5
High robotics expertise	-0.415***	-0.497***	-0.411***	-0.384***	-0.363***
	(0.0872)	(0.0879)	(0.0908)	(0.0785)	(0.0761)
High aerospace expertise	-0.402***	-0.479***	-0.381***	-0.408***	-0.386***
	(0.0901)	(0.0906)	(0.0899)	(0.0765)	(0.0870)
High robotics x High aerospace expertise					-0.0975
					(0.209)
Feasibility confidence		0.122***	0.130***	0.104***	0.104***
		(0.0368)	(0.0369)	(0.0365)	(0.0365)
Novelty confidence		0.0824**	0.0949***	0.112***	0.112***
		(0.0358)	(0.0361)	(0.0328)	(0.0327)
Age: 18-24			-0.516***	-0.512***	-0.507***
			(0.143)	(0.159)	(0.160)
Age: 25-34			-0.752***	-0.755***	-0.753***
			(0.146)	(0.150)	(0.151)
Age: 35-44			-0.776***	-0.765***	-0.761***
			(0.160)	(0.165)	(0.167)
Age: 45-54			-0.757***	-0.825***	-0.819***
			(0.184)	(0.168)	(0.170)
Age: 55-64			-1.651***	-1.664***	-1.667***
			(0.204)	(0.200)	(0.199)
USA			-0.735***	-0.780***	-0.775***
			(0.180)	(0.204)	(0.206)
Female			-0.0234	-0.0376	-0.0373
			(0.0765)	(0.0752)	(0.0752)
Bachelors			-0.204***	-0.208***	-0.209***
			(0.0703)	(0.0629)	(0.0629)
Masters			0.0235	0.0506	0.0512
			(0.0753)	(0.0798)	(0.0798)
Constant	4.836***	3.703***	4.418***	4.471***	4.469***
	(0.0318)	(0.133)	(0.189)	(0.244)	(0.245)
Solution FE	N	N	N	Y	Y
Observations	3,850	3,850	3,830	3,830	3,830
R-squared	0.023	0.045	0.067	0.210	0.210

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

Table 5. OLS Models of Feasibility Preference on Evaluator Univalent and Multivalent Expertise

	Dependent Variable: Quality Score							
VARIABLES	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8
Robotics univalent expertise	-0.0667	-0.410***	-0.0664	-0.384***	-0.417***	-0.455***	-0.421***	-0.418***
	(0.0411)	(0.0871)	(0.0411)	(0.0862)	(0.104)	(0.105)	(0.106)	(0.107)
Aerospace univalent expertise	-0.156***	-0.153***	-0.378***	-0.341***	-0.360***	-0.405***	-0.387***	-0.410***
	(0.0427)	(0.0427)	(0.0887)	(0.0881)	(0.105)	(0.107)	(0.111)	(0.117)
Feasibility score	0.576***	0.553***	0.554***	0.536***	0.531***	0.528***	0.526***	0.506***
	(0.0170)	(0.0188)	(0.0187)	(0.0199)	(0.0204)	(0.0204)	(0.0205)	(0.0213)
Novelty score	0.362***	0.374***	0.377***	0.385***	0.389***	0.386***	0.383***	0.370***
	(0.0167)	(0.0185)	(0.0183)	(0.0195)	(0.0200)	(0.0199)	(0.0200)	(0.0190)
Robotics univalent expertise x Feasibility score		0.157***		0.134***	0.178***	0.179***	0.180***	0.164***
		(0.0367)		(0.0363)	(0.0419)	(0.0419)	(0.0421)	(0.0395)
Robotics univalent expertise x Novelty score		-0.0872**		-0.0683*	-0.101**	-0.0978**	-0.0983**	-0.0826*
		(0.0370)		(0.0369)	(0.0416)	(0.0417)	(0.0418)	(0.0438)
Aerospace univalent expertise x Feasibility score			0.159***	0.137***	0.186***	0.186***	0.188***	0.172***
			(0.0365)	(0.0367)	(0.0410)	(0.0411)	(0.0417)	(0.0411)
Aerospace univalent expertise x Novelty score			-0.115***	-0.0994***	-0.139***	-0.133***	-0.136***	-0.116***
			(0.0370)	(0.0372)	(0.0414)	(0.0415)	(0.0417)	(0.0433)
Robotics univalent expertise x Aerospace univalent expertise					0.137	0.151	0.0442	0.0728
					(0.185)	(0.186)	(0.192)	(0.224)
Robotics expertise x Aerospace expertise x Feasibility score					-0.192**	-0.187**	-0.175**	-0.132
					(0.0829)	(0.0835)	(0.0841)	(0.0835)
Robotics expertise x Aerospace expertise x Novelty score					0.134	0.131	0.138	0.0816
					(0.0856)	(0.0864)	(0.0865)	(0.0814)
Feasibility confidence						-0.000564	0.000697	-0.00103
						(0.0299)	(0.0302)	(0.0301)
Novelty confidence						0.0335	0.0352	0.0435
						(0.0290)	(0.0294)	(0.0307)
Constant	0.363***	0.420***	0.405***	0.451***	0.454***	0.301***	0.324**	0.460***
	(0.0496)	(0.0548)	(0.0548)	(0.0590)	(0.0611)	(0.0857)	(0.135)	(0.107)
Evaluator covariates	N	N	N	N	N	N	Y	Y
Solution dummies	N	N	N	N	N	N	N	Y
Observations	3,850	3,850	3,850	3,850	3,850	3,850	3,830	3,830
R-squared	0.729	0.731	0.731	0.732	0.733	0.734	0.735	0.746

Note: Models 7 and 8 include the following evaluator covariates: female, USA, age range, bachelor’s degree, and master’s degree.
Robust standard errors in parentheses; *** p<0.01, ** p<0.05, * p<0.1

Table 6. Sample Comments by Topic from LDA Topic Modeling

Topic/Description	Comment Attributes	Sample Comments
Topic 1: Robotics-specific criteria Description: Comments relate primarily on the mechanics and movement of the robotic arm components.	Evaluator: A Challenge: SRA Solution number: 10 Proportion topic 1: 0.81 Evaluator expertise: Robotics univalent	- The design proposed has almost all the necessary technical details for a proper *[sic]* robotic arm, - The usage of a unique hardware elements like ARM cortex microprocessor and a uniquely designed circuit board gives a complete feasibility and flexibility for the design and utilization of this robotic arm, - Also, the usage of flexible cables to transmit electrical signals contributes to the feasibility of the overall project, - Therefore, it seems that this design can be one of the best candidates for the AstroBee robotic arm with high level of feasibility *[sic]*...
	Evaluator: B Challenge: SRA Solution number: 7 Proportion topic 1: 0.88 Evaluator expertise: Outside	I would recommend using an airflow selector valve to better control movements. 1 output only for extension of the arm and the other for the movement already described which would improve the coverage of the arm.
	Evaluator: C Challenge: PSA Solution number: 2 Proportion topic 1: 0.93 Evaluator expertise: Multivalent	The design/algorithm does not include a very important aspect of the solution, which is trajectory planning. It assumes that all the trajectories can be stored and retrieved from the database, without giving any algorithm to calculate them. Furthermore, there is doesn't tackle the problem of singularities in robotic trajectories, where the trajectory is continuous in space but not continuous in joint position/angle space.
Topic 2: Aerospace-specific criteria Description: Comments relate primarily to the extent that the design meets aerospace specific requirements.	Evaluator: D Challenge: SRA Solution number: 9 Proportion topic 1: 0.88 Evaluator expertise: Aerospace univalent	The mechanical design is well described, but the material for gears is not appropriate.
	Evaluator: E Challenge: SAM Solution number: 5 Proportion topic 1: 0.70 Evaluator expertise: Aerospace univalent	In this Design description, Gear chain Function can be seen with more than 3 gears. These 4 gears could be a point of instability or slip while exerting 20 N force as mentioned in the Functional Analysis. And any changes in a backlash between one gear may cause the gear failure in mid-space operations. So greater the number of gear mesh, the higher the chance of failure.
	Evaluator: F Challenge: SAM Solution number: 7 Proportion topic 1: 0.60 Evaluator expertise: Outside	The concept looks good, but there is not enough information given about how the SAM would react when an astronaut pulls it or what happens when it does not encounter a handrail. Lack of sensors and limited information about the materials used are also troublesome. This design is very badly described and there are multiple questions left unanswered.

Topic 3: Systems-level criteria Description: Comments focus primarily on providing an overall assessment of the design's feasibility and functionality.	Evaluator: G Challenge: SAM Solution number: 12 Proportion topic 1: 0.69 Evaluator expertise: Multivalent	This is a fantastic design with a good level of detail on the device. Including part numbers and the force plots based on the application is real work done. The system is compact and fits well within the defined parameters, mechanical drawings are also provided as per each expected configuration. Overall a well defined document.
	Evaluator: H Challenge: SPAM Solution number: 10 Proportion topic 1: 0.91 Evaluator expertise: Multivalent	The proposed design would be feasible. The Class diagram and interaction between classes show understanding of the problem and explain the detailed logic.
	Evaluator: I Challenge: SDM Solution number: 17 Proportion topic 1: 0.93 Evaluator expertise: Multivalent	The design is feasible at some point, but it needs to be more developed. The author has a good idea about the components needed and the general design, but he needs to be more specific to get a functional design.

Table 7. OLS Models of Topic Proportions (from LDA) on Univalent and Multivalent Expertise

	Topic 1: Robotics-specific criteria		Topic 2: Aerospace-specific criteria		Topic 3: System-level criteria	
VARIABLES	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Robotics univalent expertise	0.0624***	0.0672***	0.00745	0.0390*	-0.0698***	-0.106***
	(0.0208)	(0.0235)	(0.0195)	(0.0229)	(0.0201)	(0.0224)
Aerospace univalent expertise	0.00848	0.0118	-0.00952	0.0122	0.00104	-0.0241
	(0.0164)	(0.0175)	(0.0158)	(0.0174)	(0.0174)	(0.0186)
Robotics x Aerospace expertise		-0.0222		-0.145***		0.167***
		(0.0500)		(0.0401)		(0.0490)
Constant	0.279***	0.278***	0.278***	0.275***	0.443***	0.446***
	(0.00623)	(0.00628)	(0.00620)	(0.00627)	(0.00683)	(0.00690)
Observations	3,445	3,445	3,445	3,445	3,445	3,445
R-squared	0.003	0.003	0.000	0.003	0.003	0.007

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

Table 8. OLS Models of Comment Polarity on Univalent and Multivalent Expertise

	Feasibility Comment Polarity		Novelty Comment Polarity		Quality Comment Polarity	
VARIABLES	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Robotics univalent expertise	-0.0280**	-0.0317**	-0.0194	-0.0318**	-0.0499***	-0.0725***
	(0.0126)	(0.0145)	(0.0130)	(0.0152)	(0.0134)	(0.0152)
Aerospace univalent expertise	-0.00436	-0.00689	-0.0190*	-0.0275**	-0.0322***	-0.0478***
	(0.0118)	(0.0131)	(0.0112)	(0.0124)	(0.0124)	(0.0137)
Robotics x Aerospace expertise		0.0169		0.0572**		0.104***
		(0.0290)		(0.0284)		(0.0309)
Constant	0.136***	0.136***	0.137***	0.138***	0.171***	0.173***
	(0.00477)	(0.00484)	(0.00486)	(0.00493)	(0.00522)	(0.00529)
Observations	3,445	3,445	3,445	3,445	3,445	3,445
R-squared	0.001	0.001	0.001	0.002	0.005	0.007

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

Figure 1. An Illustration of Expertise Range in the Problem Area

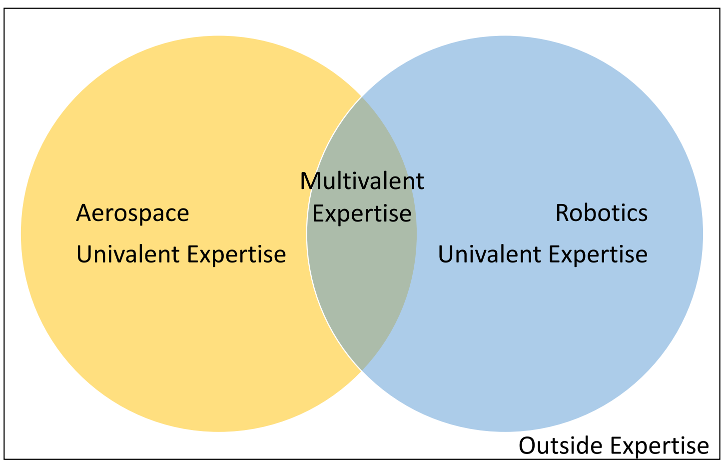
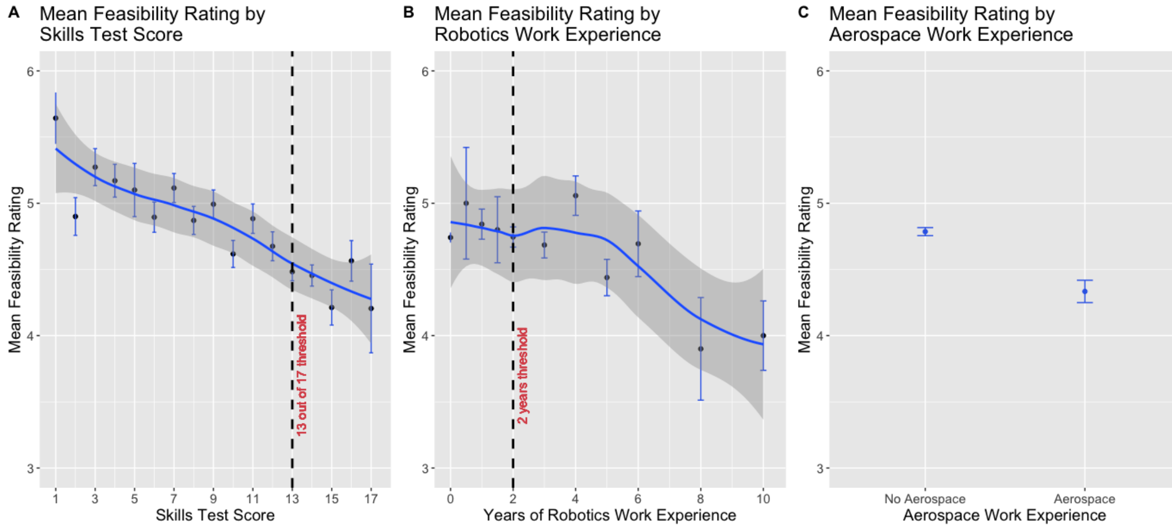
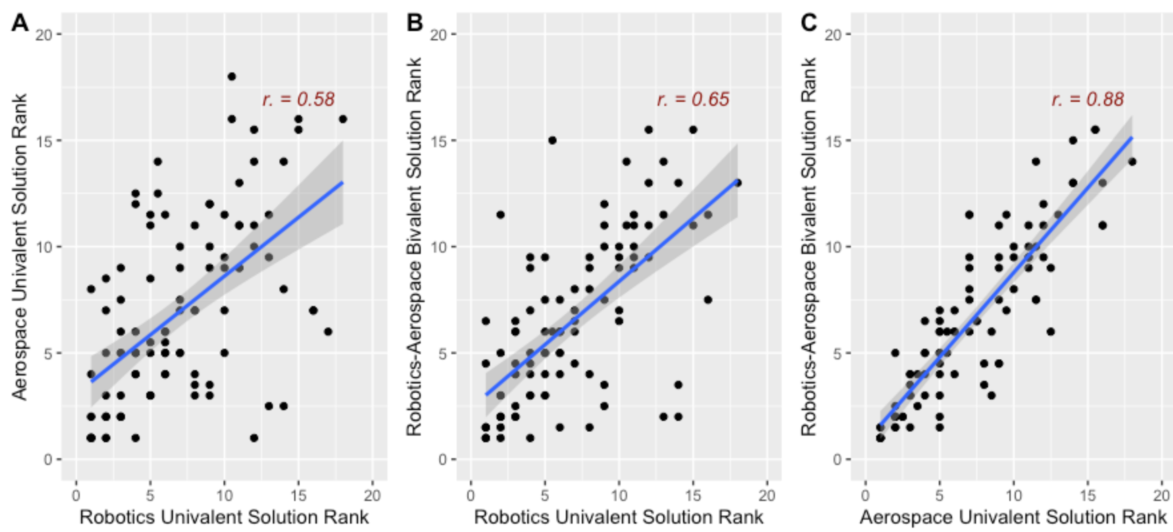


Figure 2. Distribution of Feasibility Scores by Robotics and Aerospace Univalent Expertise



Note: In Figure B, we exclude 9 evaluators with more than 10 years of work experience due to the small number of data points for computing standard errors.

Figure 3. Comparison of Challenge-Solution Rank Orders by Univalent and Multivalent Expertise Evaluators



Appendix

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

	Chi-Sq/ANOVA Test
Robotics 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 that are deleted due to missing covariate data.

Table A2. OLS Models of Topic Proportion from LDA on Univalent and Multivalent Expertise with Solution FE

	Topic 1: Robotics-specific criteria		Topic 2: Aerospace-specific criteria		Topic 3: System-level criteria	
VARIABLES	Model 1	Model 2	Model 5	Model 6	Model 3	Model 4
High robotics expertise	0.0458***	0.0503**	0.0182	0.0502***	-0.0640***	-0.101***
	(0.0172)	(0.0196)	(0.0155)	(0.0178)	(0.0197)	(0.0219)
High aerospace expertise	-0.0108	-0.00771	0.0108	0.0327**	-6.26e-05	-0.0250
	(0.0143)	(0.0154)	(0.0140)	(0.0155)	(0.0174)	(0.0187)
High robotics x High aerospace exp.		-0.0205		-0.147***		0.168***
		(0.0406)		(0.0332)		(0.0473)
Observations	3,445	3,445	3,445	3,445	3,445	3,445
Solution FE	Y	Y	Y	Y	Y	Y
R-squared	0.289	0.289	0.241	0.244	0.066	0.070

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

Table A3. OLS Regression Models of Comment Polarity on Univalent and Multivalent Expertise with Solution FE

	Feasibility Comment Polarity		Novelty Comment Polarity		Quality Comment Polarity	
VARIABLES	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
High robotics expertise	-0.0270**	-0.0328**	-0.0132	-0.0263*	-0.0470***	-0.0721***
	(0.0126)	(0.0143)	(0.0130)	(0.0152)	(0.0133)	(0.0152)
High aerospace expertise	-0.00966	-0.0136	-0.0211*	-0.0300**	-0.0364***	-0.0536***
	(0.0118)	(0.0130)	(0.0112)	(0.0123)	(0.0121)	(0.0133)
High robotics x High aerospace exp.		0.0265		0.0599**		0.115***
		(0.0295)		(0.0279)		(0.0303)
Observations	3,445	3,445	3,445	3,445	3,445	3,445
Solution FE	Y	Y	Y	Y	Y	Y
R-squared	0.060	0.060	0.045	0.046	0.071	0.073

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

Figure A1. Screenshots of Evaluation Procedures

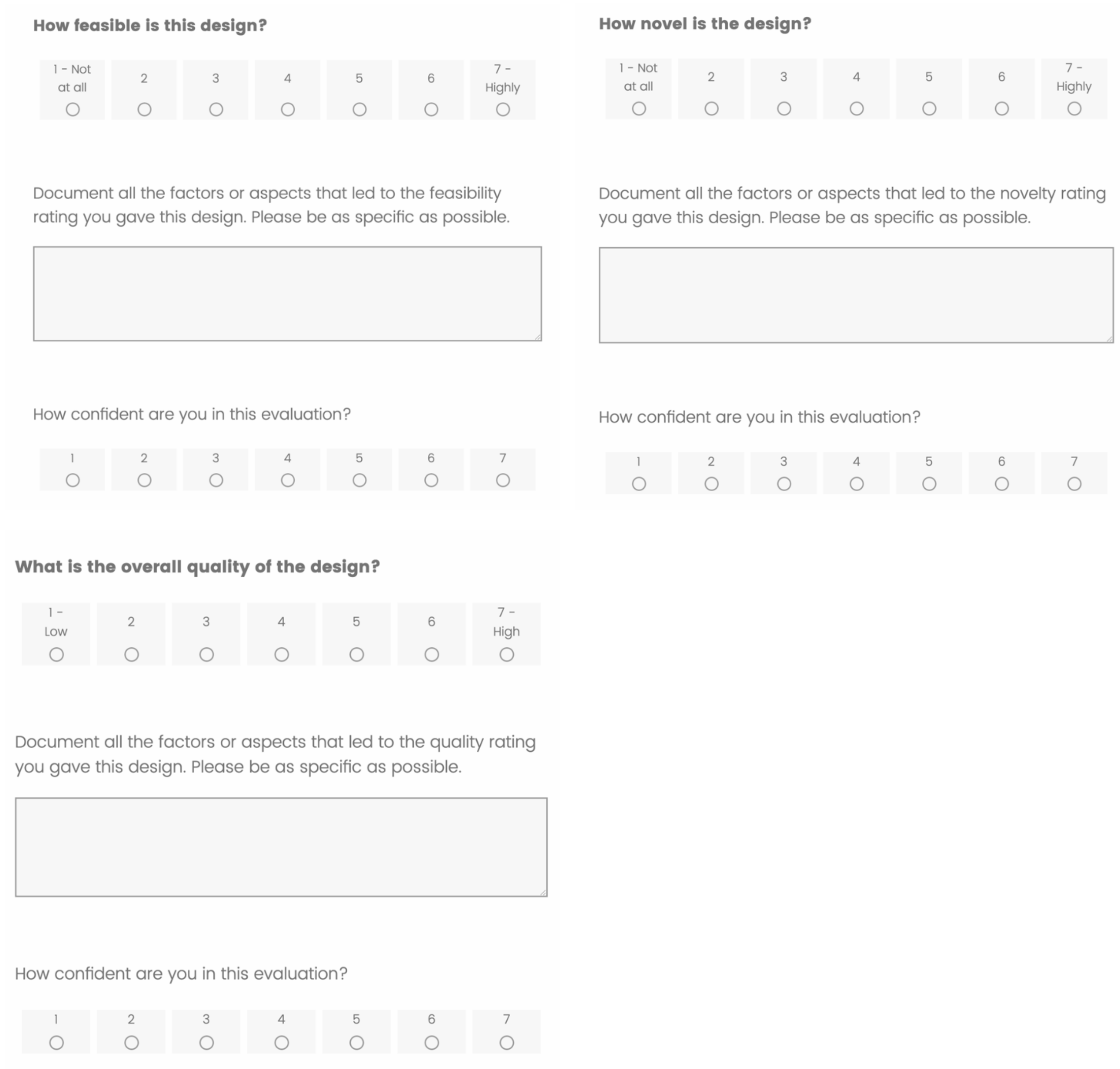
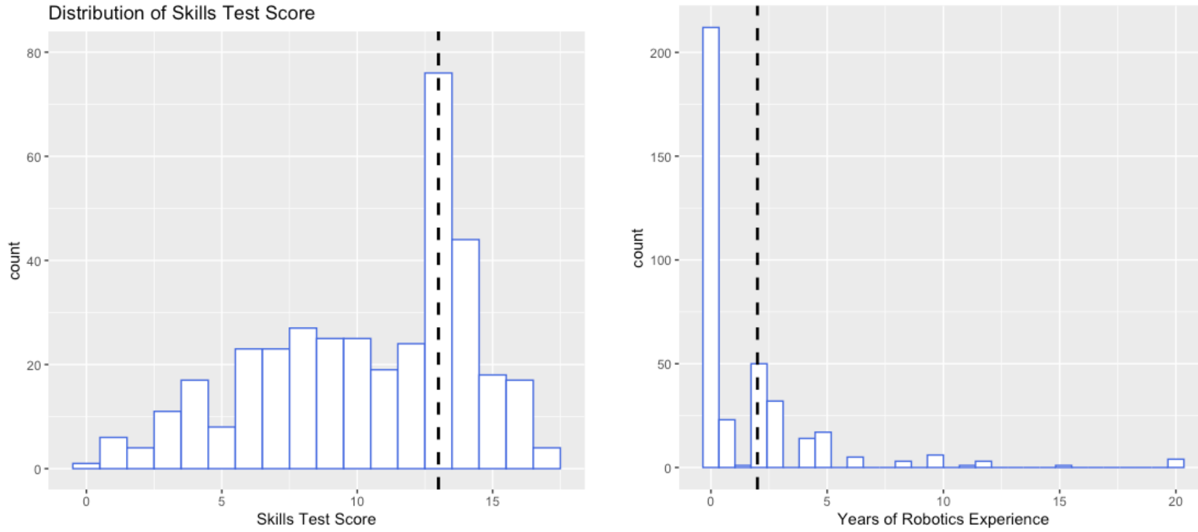


Figure A2. Distribution of years of robotics/mechatronics engineering work experience



Note: Black dashed line corresponds to the threshold score on the skills test (13 or more) and years of work experience in robotics/mechatronics (2 or more) for univalent expertise in robotics.