Working Paper 23-014

Scoring and Funding Breakthrough Ideas: Evidence from a Global Pharmaceutical Company

Joshua Krieger
Ramana Nanda
Ian Hunt
Aimee Reynolds
Peter Tarsa

Harvard Business School

Scoring and Funding Breakthrough Ideas: Evidence from a Global Pharmaceutical Company

Joshua Krieger
Harvard Business School

Ramana Nanda
Imperial College London

Ian Hunt
NBIR

Aimee Reynolds
NBIR

Peter Tarsa
NBIR

Working Paper 23-014

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

Scoring and Funding Breakthrough Ideas: Evidence from a Global Pharmaceutical Company*

Joshua Krieger
Harvard Business School

Ramana Nanda
Imperial College London

Ian Hunt
NIBR†

Aimee Reynolds
NIBR†

Peter Tarsa
NIBR†

November 21, 2023

Abstract

We study resource allocation to early-stage ideas at an internal startup program of one the largest pharmaceutical firms in the world. Our research design enables us to elicit every evaluator's scores across five different attributes, before seeing how they would allocate capital to the projects in a head-to-head comparison. In head-to-head comparisons, evaluators displayed a systematically higher preference for projects that scored high on execution-related attributes, compared to the organization's proposed weight on these attributes. Because of this, projects of similar overall quality perceived as being high risk and high reward were systematically penalized relative to projects perceived as less transformational, but safer bets. Our results shed light on a potential mechanism for why breakthrough ideas are handicapped in R&D funding.

JEL Classifications: O31, O32, Q55, L65

Keywords: project selection, research and development, pharmaceuticals, financing innovation

*We are grateful to John Galvin and Karin Parodi for research assistance with the live surveys. We also thank Kevin Boudreau (discussant), Jacqueline Lane, Misha Teplitskiy, and participants at the 2021 Harvard/Duke/UCSD Workshop on Strategy, Innovation, and Entrepreneurship, the HBS Entrepreneurship Unit brownbag lunch, and the AEA 2022 annual meetings for helpful comments. This paper is part of a project that has received funding from the European Research Council (ERC) Horizon 2020 research and innovation programme, Grant Number 865127.

†Novartis Institutes for BioMedical Research

1 Introduction

Resource allocation decisions by research and development (R&D) funders can play a critical role in determining which among different potential transformational innovations get commercialized. Given the great uncertainty under which such resource allocation decisions are made and the outsized role that relatively few gatekeepers play in funding early stage innovation, there is growing interest in understanding decision-making by these gatekeepers, and the degree to which the *evaluation process* might systematically impact the direction of R&D. For example, proposals to reform grant-making processes with "golden ticket" awards and even lottery systems have gained traction and some high profile experiments with the largest government science funders (Chawla Chawla; Adam et al. 2019). The White House Office of Science and Technology Policy's FY2025 budget priorities memo includes a call to further "experiment with funding processes to better achieve agency R&D missions."[1]

Most of the academic and policy work on the structure and process of R&D funding has focused on research-related grant-making by foundations or agencies within government, (Boudreau et al. 2016; Azoulay and Li 2021; Lane et al. 2021; Franzoni et al. 2022). We know much less about the evaluation processes involved in the development stage of R&D funding. A large share of the transition from research to development is undertaken by the internal R&D teams of large private sector organizations. In such R&D environments, industry scientists and executives judge the merits and opportunities of uncertain R&D projects that shape the future. Firms charge these evaluators with the important task of translating the firm's corporate goals into a portfolio of projects that reflects firm-level strategy. Yet, we know little about how the gatekeepers at these institutions think about resource allocation decisions, and in particular the financing of early-stage innovation, which tends to be concentrated in fewer number of bold bets compared to research funding. For

[1] See https://www.whitehouse.gov/wp-content/uploads/2023/08/FY2025-OMB-OSTP-RD-Budget-Priorities-Memo.pdf

example, do different scoring and evaluation regimes have the potential to systematically bias such internal R&D funding towards different types of projects and thereby distort the variety of projects that win funding?

The dearth of research on large firm R&D funding is driven in large part by two big measurement challenges associated with studying any resource allocation in the context of R&D funding in large firms. First, one must get access to the full risk-set of projects actually considered for investment, rather than just the final set of projects actually developed. This is typically much harder in an environment with commercially sensitive information compared to examining research funding by grant making bodies. Second, researchers need to observe how the same individuals score the same projects under different evaluation regimes. Without this, any observed differences in the outcomes of scoring regimes can be confounded by different preferences of the specific individuals involved in making the assessments.

We overcome these challenges by using unique data on actual projects in the real world setting of an internal startup program at Novartis, one of the largest pharmaceutical companies in the world. The program, called Genesis Labs, was set up in order to deliver "transformative, breakthrough innovation" by offering a fast-paced funding and development process that ran parallel to the typical project selection and funding paths.[2] After soliciting applications from teams across the organization, a three step peer-review screening process pared 165 applications down to a shortlist of twelve. The finalists then pitched to a committee of scientific experts responsible for selecting four to fund. This final stage is the context of our study.

Three features of the setting are particularly useful for our analysis and provide a unique window into decision making by the R&D experts who propose and select innovative projects at Novartis. First, R&D personnel from across the organization were invited to view the

[2] Edmondson and Gulati (2021) describes the formation of the Genesis Labs initiative, which set out to "catalyze more-radical innovation" by encouraging cross-disciplinary teams to propose their "dream projects" as part of the competitive funding contest.

pitches and participate in a live survey soliciting their evaluation of projects. These employees had strong and relevant technical backgrounds for understanding and evaluating these ideas as well deep knowledge of the organization (average tenure at the firm was 11 years). More than two-thirds of those who participated in the study had MDs and/or PhDs and several had been involved in screening applications for resource allocation decisions in similar programs in prior years. We collected data from 141 participants, resulting in 503 participant-project evaluation pairs across the 12 projects.[3] This sample size of individuals—who were not on the actual panel but very representative of those who were—provides us with substantially greater power than is usually possible when studying decision making in a real-world setting, as committees are usually much smaller.

Second, restrictions stemming from COVID-19 necessitated that these pitches all had to be done remotely, so the virtual setting for the presentations meant that participants saw exactly the same information as the actual evaluation committee. The real-time evaluations picked up each individual's beliefs without any confounding factors arising from discussing (or overhearing discussions) of the pitch with other participants.

Third, we asked participants to "live" score projects separately along each of five dimensions the organization has chosen as their evaluation criteria, then later allocate hypothetical dollars in head-to-head allocations across the projects in each of three sessions. This two step measurement in the research design enables us to first elicit perceived strengths and weaknesses for each project across different dimensions the organization had noted it cared about and analyze how the *same individual's* resource allocation decisions map to their independent project-specific attribute evaluations. This "within-person" design addressees important sources of unobserved heterogeneity that are typically hard to address in real-world studies of decision making by experts. It also allows us to examine whether the implicit

[3] Projects were grouped into three sessions based on their broad thematic focus and most participants joined for one of the sessions.

weights assigned to the criterion by the evaluators in head-to-head allocations deviated in systematic ways from the proposed weights the organization had explicitly articulated it wanted to use.

We find that the average scores received by projects were closely clustered—the very top and bottom ranked projects only differed by about 20% (4.0/5 vs. 3.3/5)—but the variation in scores across the different attributes received by the same evaluator ("attribute imbalance"), was substantial. Most notably, we observe a common tradeoff between attributes related to the project's potential impact and those related to execution risk. Even after controlling for average overall score, we find a persistent negative correlation between impact and execution attributes. It is worth emphasizing that these were the top twelve of 165 proposals, so low execution scores should not be seen as reflective of an absolute issue with execution, but rather a relative difference compared to other projects that also reflects for example, the timescale within which impact can be realized.

Second, we compare evaluators' project attribute scoring (Survey #1) to their head-to-head dollar allocations across projects (Survey #2). We find that the head-to-head comparisons are associated with stronger preferences across projects, leading to a "pulling apart" of project scores. Specifically, we measure much greater separation between project allocations than would have been suggested by their Survey #1 attribute scores. However, this greater conviction and related separation appears to come at a cost: in head to head allocations, evaluators can impose their subjective weights on the importance of different attributes, which we find to be systematically different from the organization's proposed goals. In our context, we find that despite the program's efforts to signal their desire to fund higher risk/reward projects, the scientist evaluators weight execution more highly, and all else equal, penalize the combinations of project attributes associated with high risk and high potential impact in favor of those that are safer bets. This bias in preference for safer bets relative to the organization is large enough to reorder the projects that would have been funded had the organization used its own weights.

Our results provide a potential mechanism for a perceived lack of funding for the most transformational ideas: systematically losing out in head-to-head resource allocation decisions to ideas more balanced in their attributes, even when the latter are perceived by evaluators as having less potential for impact. This pattern is even more striking because we find it in the context of a company program that was quite explicitly seeking novel and high impact projects. Though a heuristic that tends to favor consistent within-project qualities may be rational and even advisable in certain contexts, recent evidence suggests it could be sub-optimal in contexts such as early stage innovation (Malenko et al. 2021). This is because even the most breakthrough concepts—as recognized in early stage venture capital—may have aspects that are not yet on solid footing. The degree to which evaluators focus on these potential flaws (that may be addressable), rather than the potential upside and the option value of high risk/high reward investments, influences the degree to which the most breakthrough ideas have a chance to evolve into something more impactful.

Our work is related to several strands of the literature. Prior work looking at the pharmaceutical industry has highlighted distortions that stem from *external forces* such as intellectual property, competition and financing that limit innovative novelty and R&D exploration. The limited monopoly period granted by patents pushes firms to work on technologies with shorter paths to market (Acemoglu 2012). For example, drug developers are more likely to invest in drug candidates targeting diseases that have shorter clinical trials due to surrogate endpoints (Budish et al. 2015). Firms may also herd into similar technological areas to take advantage of information spillovers across competitors (Bloom et al. 2013; Bryan and Lemus 2017; Krieger 2021). Risk aversion and capital constraints have also been shown to limit investment in novel technology (Nanda and Rhodes-Kropf 2013, 2017; Krieger et al. 2021). Our paper instead focuses on the *internal* forces from information acquisition and aggregation. We show how even in the presence of organizational will to invest in novelty, the (multi-dimensional) evaluation of project quality has the potential to inadvertently stymie the funding of high-risk, high-reward projects if evaluators are given the discretion to impose their own weights on these attributes when making resource allocation decisions.

Our work also speaks to the nascent but growing literature on project selection in science funding, which has largely focused on research grant processes (often at the National Institutes of Health) and on the composition of the selection committees and applicants (Boudreau et al. 2016; Gallo et al. 2016; Li 2017; Myers 2020; Azoulay and Li 2020; Lane et al. 2021). Our resource allocation results are in line with what Lee (2015) calls "commensuration bias" in peer review, where evaluators systematically prioritize review criteria based on their own personal values and category weights, rather than the organization's priorities. Further downstream, the selection of startups by venture capitalists has also been an area of research interest (Kerr et al. 2014; Lerner and Nanda 2020; Ewens et al. 2018), though that stream of work has mainly focused on the consequences of receiving private capital, and on distortions in the willingness of venture capitalists to invest in different areas. As far as we know, this is the first study to examine drug development resource allocation in a "live" real-world setting.[4] Other studies have studied pharmaceutical firm's investment allocation using retrospective analysis of funded projects and their outcomes (Cook et al. 2014; Morgan et al. 2018; Shih et al. 2018).

While a study on a single organization creates legitimate concerns about the generalizability of the findings, we believe that our research identifies an important core tradeoff between highly structured R&D proposal scoring systems and head-to-head comparisons. In the former, an organization can impose their explicit weights and priorities, but may find a lack of separation between projects in their overall scores, as experts are not forced to compare the bundle of attributes relative to a benchmark. Head-to-head comparisons reveal stronger

[4] Notably, other recent papers have studied R&D project selection in experimental settings (Carson et al. 2022), and how prize structures influence open innovation competitions hosted by firms (Graff Zivin and Lyons 2021).

preferences, but allow individuals to impose their own values on the selection process. Those personal R&D preferences can be strong enough to reorder project rankings and substantially change what projects firms pursue. Since our research setting is a program that intensely signaled the firm leadership's preferences for certain types of project characteristics, we believe our results are relevant to equivalent internal R&D programs in other firms. Given the millions of dollars collectively channeled in translation and development by large R&D-driven firms each year, we believe these results have important implications for how such programs are designed and the potential training involved for those engaged as gatekeepers in these key resource allocation decisions.

2 Research Setting & Design

2.1 Novartis Genesis Labs Program

Our research setting is an internal project funding process at Novartis. Our data comes from the third edition of this Genesis Labs program, which the company ran annually starting in 2017. The company initiated this program to identify and fund "transformative breakthrough innovation" in parallel to the existing processes for selecting innovation within the R&D organization. The Genesis program was initiated to nurture high risk, high reward projects that might otherwise have harder time gaining support in the company's traditional R&D structure. By funding radical projects and pulling their teams out of their usual R&D roles and into an "internal startup" environment, the leadership's aim was to "foster a culture of entrepreneurial thinking to drive cross-disciplinary collaboration." Genesis Labs aimed to be a proving ground for such projects—bearing the higher risk of the project, while also providing pathways back into those home R&D groups (after 18 months and significant de-risking), so that the company could fully realize the high potential of the surviving ideas.[5]

The program solicited ideas from R&D teams across more than 20 global research offices and four different R&D divisions. Teams generally consisted of 3–5 scientists, often with cross-functional backgrounds and representing multiple offices. The winning teams received a budget of roughly $2 million in special funding to support 18 months of de-risking their project. Winning teams were assigned expert mentors, and received access to centralized business support services like legal, information technology, operations, etc.. The Online Supplement Section A provides additional information about Genesis Labs program, including the project selection process and the categories for evaluation.

Initial filtering took the applicant pool from 165 to 51 projects. The remaining projects underwent a round of peer review, in which four to seven Novartis scientists scored each project. The Genesis labs "core team" (roughly 25 internal scientists) aggregated the peer review scores and selected the 12 finalist projects for Pitch Day.

On Pitch Day, a panel of eight company leaders and one external academic expert ultimately determined the winners. The organizers promoted the event as a celebration of novel research at the company and invited all R&D employees to attend and participate in project scoring. Pitch Day took place in August 2020. Due to the COVID-19 pandemic, all the pitches were conducted virtually, with the presenters, panelists and audience all participating remotely through video conferencing software. The virtual nature of the event was an important feature of our study, because it both enabled a broader group of company scientists to participate as evaluators, and enabled us to capture their judgment in a way that was not directly impacted by their peers (e.g., crowd reactions, body language).

[5] Over the first three Genesis Labs funding cycles, the program funded 15 projects. As of April 2022, 11 projects had completed (4 ongoing), and 7 had "progressed" into either enabling technologies or methodologies used in Novartis's drug development R&D teams. The program estimated that 37 R&D projects across the company had been impacted by these Genesis Labs funded ideas. On the human capital side, 119 R&D associates have participated on the funded projects, experiencing over 30 promotions and/or career advancements in the five years since the program's inception.

The organizers invited attendees in the pre-event emails and with links and QR codes shared at the start of each session. The pitches were organized into three sessions consisting of four projects each (grouped by broad topics). Teams had 10 minutes to present, followed by 10 minutes of questions and answers with the panelists. In between sessions, there was a 15 minute break.

2.2 Research Design

Our main analyses involve the evaluation of the final 12 projects presented during the Pitch Day event. A critical element of our research design is that we are able to first elicit scientists' views on particular project attributes, and to then observe the *same evaluators'* resource allocation decisions. This design provide a unique window into the (typically unobserved) decision-making process that underlies resource allocation and ensures that any differences we observe between two scoring regimes is not confounded by differences in the preferences of the individuals involved in the evaluation.

A cost of doing this two-step evaluation design is that it puts a greater burden on inclusion criteria in some of our analyses. Table A.1 summarizes the characteristics of participants , while Table A.2 shows that the subsample of individuals who completed both steps of scoring does not differ substantially.

The two-step sequence of session events and project evaluation are depicted in Appendix Figure A.1, along with examples of the survey interface. Appendix Section A.1 describes the scoring process.

3 Results

3.1 Finalist and Participant Characteristics

Before analyzing the Pitch Day evaluation sample, we assess how our participant evaluators compare to other expert evaluators within the firm. We find that the Pitch Day participant evaluators gave scores that aligned with the prior round of semifinalist peer review. Figure A.2 in the Supplement describes the peer review scoring differences between the 12 selected finalists and the 39 non-selected semifinalists. Finalists scored higher across all categories in the peer review round (Panel C), though their normalized distribution suggests that the level of inter-reviewer disagreement in average weighted scores was similar for projects selected vs. not selected (Panel B). Figure A.3 shows a positive correlation between the average weighted scores from peer review and the Survey #1 Pitch Day audience weighted scores. Three of the top four projects from peer review scoring also were in the top four of the Survey #1 results.

Table A.1 of the Online Supplement provides additional descriptive statistics of the final participant sample and their self-reported expertise.

3.2 Evaluation of Project Attributes

Our first set of findings describes the independent scoring outcomes and their scoring distributions *across* evaluations. While the head-to-head dollar allocations of Survey #2 more closely reflect a typical portfolio selection committee meeting, the Survey #1 scores provides a view into how individuals assess project-specific qualities and the relative strengths across those attributes.

Figure 1, shows the Pitch Day Survey #1 scoring results, by project rank. Two features immediately jump out. First, the figure shows that three of the top four average weighted score projects were selected by the panel for funding. This general consensus is reassuring in that all three sets of evaluators—peer reviewers, pitch day evaluation participants, and selection panel—overlapped in how they judged quality (overall).

Second, we see that very little separates the project's average weighted scores. The top ranked project has an average score of 4.0, while the bottom ranked project has an average of 3.3, and the top six projects all have averages within 0.5 of one another. This clustering of average scores is not surprising, given that the 12 projects had already survived multiple rounds of peer review. Third, we see quite a bit of heterogeneity for any given project's scores, such that the 90% confidence intervals average more than two full Likert scale points, and all the average scores fall within the 90% confidence intervals of the other projects. The one funded project outside of the top four was 9th in the average audience scoring; however, it ranked third in audience average Transformational score—suggesting that the panelists might have put an additional premium on that attribute.

In any setting with Likert-scoring a concern is that evaluators apply different standards, resulting in a potentially unbalanced mix of "generous" and "harsh" evaluator scores. Panels A of Figure A.8 in the Supplement shows our estimate of individual participant generosity based on evaluator fixed effects. Though a small group of evaluators are indeed harsher or more generous than the median evaluator, controlling for reviewer generosity leads to only minor changes in project's aggregate scores or their rank order (Panel B).

A feature of our participant sample is that the large majority would be qualify as "scientific experts" by any casual definition. However, an individual's research focus is often quite specialized, even within drug development. For that reason, we asked each participant to rank their expertise *relative to the particular pitch project*, and participants responses were normally distributed (see Appendix Figure A.7, Panel B). We find that "expert" evaluators—those responding with a 4 or a 5 in project-specific expertise—tended to be slightly more generous. Scores were slightly higher among experts, even after controlling for participant and project fixed effects (see Panel B of Figure A.7 and Table A.3). Similarly, we found evidence of a small "home team bias," by which evaluators were slightly more generous in scoring projects that had at least one team member from the evaluator's Novartis division. Though experts and proximate colleagues appeared to score projects more favorably, tenure at the firm was significantly associated with harsher scores (see Column 4 of Table A.3). Overall, evaluator characteristics appear to have a significant but small (in magnitude) impact on project aggregate scores.

3.3 Execution vs. Impact Attribute Tradeoffs

In our second set of results, we examine category score heterogeneity *within* evaluations. We find that certain attributes and certain pairs of attributes are more likely to be associated with overall within-evaluation imbalance.

We find substantial across attribute variation within project evaluations. Some projects are especially prone to such incongruous attribute evaluations. Figure A.4 in the Supplement summarizes the within-evaluation (across attribute) score range. 50% of evaluations have a gap of at least two points between their top and bottom attribute scores (Panel A). However, we do not find a correlation between attribute inconsistency and average project scores (Panel B). In other words, some projects were consistently inconsistent across attributes, and those might end up anywhere in the average weighted score distribution.

Figure 2 shows the pattern driving attribute inconsistency: a tradeoff between impact and execution scores. Even after controlling for overall project quality bins (terciles of average weighted score), we find a negative corellation between Transformational Potential and Path-to-Execution scores or between the total impact scores Transformational Potential + Breadth) and total execution scores (Path-to-Execution + Timescale). This tradeoff is not merely mechanical, as supplementary analyses show a positive or non-existant corellation between attributes within the execution and impact categories (see Figure A.5) and that within-evaluation attribute imbalance is most likely due to high Transformational Potential scores and low Path-to-Execution and/or Timescale scores (Figure A.6). In Figure A.9, we normalize attribute scores and again find that attribute imbalance is most commonly due to differences in attributes that capture probability of success and those that track potential impact.

3.4 Resource Allocation Results

Our third set of findings shows how relative project evaluations shift under a head-to-head comparison. Overall, we find that the head-to-head comparisons amplify relative project preferences, reorder relative ranking, and penalize attribute inconsistency.

Comparing Predicted vs. Actual Allocations. The advantage of our data gathering process is that we see how evaluators' independent project scores compare to their dollar allocations in the end-of-session survey. To quantify the difference between resource allocation choices and what we might expect from project attribute scores, we calculate each participant i's *predicted allocation* for project j based on ij's Survey #1 weighted score relative to their other (same session) scores:

$$Predicted\hat{A}llocation = \frac{WScore_{i,j}}{\sum W.Score_i} \times 100$$

Thus, if an evaluator had given all four projects the same total weighted score, then the predicted allocation for each would be \$25. If the four projects had weighted scores of 4.5, 3.5, 3, and 2, then they would have predicted allocations of \$33.33, \$25.93, \$22.22, and \$18.52, respectively.

Reasons for deviation. Deviations from the predicted allocation might fall under three general explanations. The first is timing and misremembering: with a lot of information presented in a short period, evaluators might inaccurately remember their project-specific evaluations as they approach the end-of-session survey. We believe this explanation to be highly unlikely in our research design, because we reminded all participants of their own project weighted scores by displaying their Survey #1 outcomes above the portfolio allocation sliders. We also control for project presentation order in our allocation regressions (see Column 4 of Table A.4), and our findings hold.

The second reason relates to evaluation scales and conviction. An evaluator might have clear rank order preferences over projects, but still "politely" score them closely to one another in Survey #1 (e.g., 4.0, 3.9, 3.8, 3.7), even though the evaluator really only wanted to fund the first two projects. If that were the case, we'd expect the Survey #2 allocations to amplify, but preserve, rank order differences from Survey #1.[6]

Deviations might also reflect individuals' relative preferences for certain "bundles" of project attributes. Unlike the independent category scoring, the end-of-session allocation gives participants a chance to assert their own values over the relative importance of certain attributes or combinations of attributes. Combined with our data on evaluators' independent attribute scoring, the Survey #2 allocations reveal additional signals about which *types* of projects the scientists prefer. Empirically, these types of deviations would show up in the allocation data as both changes in project rank and as penalties/premiums for projects with particular combinations of attributes.

3.5 Actual vs. Predicted Allocations

We find evidence of both amplification and reordering of project rank preferences. Panel A of both Figure 3 and Figure A.10 shows how predicted allocations (based on Survey #1 weighted scores) fall into a fairly narrow distribution, with the large majority of evaluator-projects expected to get between \$20 and \$30. However, that actual allocation distribution is much wider, with an interquartile range of \$4.5–\$43.5 (mean=\$25). This "pulling apart"

[6] This amplification need not be symmetric or linear, but those relationships are easily tested both graphically and in regression analyses, by looking at the relationship between predicted and actual allocation for different levels of the predicted allocation distribution. Indeed, we find a very linear relationship between predicted and actual allocations which is symmetric about the median.

of the predicted allocation is most evident when we graph actual allocations by predicted allocations (see Figure 3, Panel A). The relationship is linear, with a slope greater than 1, and the pattern is symmetric, such that projects below the median see allocations negatively affected by roughly the same proportions as those above the median benefit.[7] Our regression analysis in Table A.4 shows that on average, a $1 increase in predicted allocation translates into a $4.49 increase in actual allocation (Column 3). That affect decreases slightly to $2.54 when we control for the participant-project's Survey #1 rank order and session presentation order (Column 4).

Stronger preferences do not necessarily mean different relative rankings. However, we indeed find project preference reordering from Survey #1 to Survey #2 (see Figure 3, Panel B). 48% of participant-projects evaluations undergo a change of one or more in rank ordering (e.g., moving from 4th to 3rd ranked project within the given session), and 14% move more than one full rank (e.g., 1st to 3rd).[8] These rank changes suggest that not only does the portfolio allocation process amplify (i.e., "pull apart") project scores based on their rank and conviction, but it also reveals different signals about projects' relative value.

We also find that the expertise effects are further exaggerated in the portfolio allocation choices. As described in Section 3.2, expert and home-team effects are already reflected in the independent weighted scores, and therefore in the predicted allocations. Our regressions in Table A.4 (Columns 1–4) suggest that experts boost allocations another $4.7–$7.7 dollars beyond what their Survey #1 scores predicted, and that "within participant expertise"

[7] Figure A.11, Panel A plots the relationship between quantiles of predicted allocation and the difference between actual and predicted allocation, after controlling for a number of evaluation characteristics (evaluator expertise, home-organization effects, attribute inconsistency). The results show a striking positive and highly linear relationship between predicted allocation and the gap between actual and predicted allocations. Regressions analyses (not shown, for brevity) confirm no statistically significant difference in slope for projects below and above the median in predicted allocations.

[8] A potential confounder for rank reordering is updating due to the peer score "treatment" that half of Survey #2 respondents randomly received. When we limit the sample to only those who were in the control group (no peer information), we see lesser, yet still substantial, degree of rank updating. 34% of participant projects move one or more ranks. And 10% move more than one full rank.

(participant's project expertise relative to their total expertise on all projects in that session) is also significantly associated with greater dollar allocations (Columns 6 and 7).

However, we find that within-evaluation attribute inconsistency is associated with an allocation penalty. Even after controlling for predicted allocation, expertise, Survey #1 rank and presentation order, peer score treatment, and project fixed effects, attribute inconsistency is penalized in the Survey #2 allocations.

3.6 High Risk, High Impact Projects Penalized

Finally, we show how the trade-off between execution and impact attributes drives the stronger preferences ("pulling apart") and rank changes from Survey #1 to Survey #2.

As discussed in Section 3.2, within-evaluation attribute scores varied substantially, and that attribute inconsistency is most often due to divergent execution and impact scores. Next, we show how that attribute trade-off translates into a premium/penalty in the resource allocation stage.

Figure 3 shows how that trade-off distorts final (constrained) dollar allocations. The binscatter plot in Figure A shows that after controlling for that evaluator's relative ranking of the project (based on Survey #1 weighted scores), peer score treatment status, and evaluator self-reported expertise, evaluators impose an allocation penalty when impact scores (Transformational Potential + Breadth) exceed execution scores (Path-to-Execution + Timescale) by two or more. When evaluators gave roughly equal impact and execution scores, or in the more rare case when execution scores exceeded Impact scores, they gave the project a boost in Survey #2 allocations.

We also look at how specific pairwise attribute differences are penalized or rewarded in the constrained portfolio allocations. Figure A.12 in the Supplement shows the relationships between each attribute pair, starting with the four most frequently divergent pairs (Transformational–Path-to-Execution, and Path-to-Execution–Breadth). We clearly see a premium put on Path-to-Execution and Timescale, while we see a penalty for projects that are high in Transformational Potential and Breadth, but lower in feasibility.[9]

We confirm these relationships in the regression analyses reported in Table A.5. We find that projects with feasibility / path to execution as their best attribute receive an additional allocation premium: \$6.02 more than projects with Timescale as their greatest strength, and \$7.42 more than projects with Transformational Potential as strongest category. After controlling for predicted dollar allocation, peer treatment status, and Survey #1 weighted score rank, we see clear trade-offs between pairs of project attributes (Columns 3–7). The largest trade-off for the pairs was Transformational Potential vs. Timescale. Each point of difference between those two attributes is associated with a \$2.37 average change in additional dollars allocated ($p < 0.05$) after controlling for projected allocations. Aggregating across categories, we find a strong ($p < 0.01$) association between impact–execution scores difference and dollar allocation. Every additional point difference between impact and execution scores is associated with a \$1.73 allocation penalty (Column 8).

Last, these allocation penalties do not merely strengthen convictions and move a few dollars at the margin, they also change project's relative rankings. Panel B of Figure 3 mimics Panel A, except the outcome is the evaluator-project's relative rank change (from those implied by their Survey #1 scores to their ordering in Survey #2). Again, we see a strong negative correlation with Impact–Execution scores. Projects with high (low) Impact scores and low (high) Execution scores are more likely to move relative ranks. Participants put more emphasis on execution, especially when those attributes diverge from impact potential.

In summary, our results show that a specific flavor of attribute divergence drives the separation in project preferences we see going from Survey #1's independent scoring to the constrained allocation in Survey #2. Evaluator's portfolio allocations signal a distaste for

[9] The binscatter plots adjust for Survey #1 evaluator-project rank (within the session), evaluator expertise, "home team" bias, and peer score treatment status.

novelty and breadth when those characteristics comes at a cost of increased execution risk, and they often do. Since the combination of high Transformational Potential and low Timescale / Path-to-Execution scores are the most common pairs driving attribute inconsistency within evaluations, that distaste mostly impedes highly Transformational Projects and rewards less risky ones, even though the program intended to encourage the opposite prioritization.

4 Discussion & Conclusion

Using data on the scoring of actual projects at an internal startup program at one of the largest pharmaceutical companies in the world, we provide a unique window into R&D funding decisions within large companies. Our research design enables us to first elicit perceived strengths and weaknesses for each project across different dimensions and analyze how the *same individual's* resource allocation decisions map to their independent project-specific attribute evaluations. Our sample of 141 scientific experts—who were not on the actual panel but very representative of those who were—provides us with substantially greater power than is usually possible when studying decision making in a real-world setting.

We find that the R&D professionals in our sample are prone to see a tradeoff between execution and impact attributes, even when their independent project evaluations show minimal differences in project average scores. In the head-to-head allocation decisions, we see a "pulling apart" of relative project scores that both amplifies the independent project assessment differences and often reorders project rank. A major driver of the reordering is evaluators penalizing or promoting projects with a gap between their execution and impact attributes. That scientist evaluators display such preferences in the context of a competition that was *explicitly* celebrating novelty and high risk, high reward projects, suggests that the norms of scientific skepticism might inhibit high-variance exploration, even when the organization signals an appetite to pursue such projects. If this pattern is present in our research setting, then it is likely such“organized skepticism” is even more prevalent in public R&D agencies’ funding committees and in other private sector R&D organizations less focused on funding novel explorations.

Our study also highlights a tradeoff between highly structured R&D proposal scoring systems and head-to-head comparisons. In the former, the organization can impose their explicit weights and priorities, but may find a lack of separation between projects in their overall scores as experts are not forced to compare the bundle of attributes relative to a benchmark. Head-to-head comparisons (as in Survey #2) get around this and reveals stronger preferences, but allows individuals to impost their own values on the selection process. Those personal R&D preferences can be strong enough to reorder project rankings and substantially change what projects firms pursue.

References

Acemoglu, D. (2012). *Diversity and Technological Progress*, pp. 319–360. University of Chicago Press.

Adam, D. et al. (2019). Science funders gamble on grant lotteries. *Nature 575*(7785), 574–575.

Azoulay, P. and D. Li (2020, March). Scientific grant funding. Working Paper 26889, National Bureau of Economic Research.

Azoulay, P. and D. Li (2021). *4 Scientific Grant Funding*, pp. 117–150. Chicago: University of Chicago Press.

Bloom, N., M. Schankerman, and J. Van Reenen (2013). Identifying technology spillovers and product market rivalry. *Econometrica 81*(4), 1347–1393.

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

Bryan, K. A. and J. Lemus (2017). The direction of innovation. *Journal of Economic Theory 172*, 247–272.

Budish, E., B. N. Roin, and H. Williams (2015, July). Do firms underinvest in long-term research? evidence from cancer clinical trials. *American Economic Review 105*(7), 2044–85.

Carson, R. T., J. Graff Zivin, J. J. Louviere, S. Sadoff, and J. G. Shrader (2022). The risk of caution: Evidence from an experiment. *Management Science*.

Chawla, D. S. 'golden tickets' on the cards for nsf grant reviewers. *Nature*.

Cook, D., D. Brown, R. Alexander, R. March, P. Morgan, G. Satterthwaite, and M. N. Pangalos (2014). Lessons learned from the fate of astrazeneca's drug pipeline: a five-dimensional framework. *Nature reviews Drug discovery 13*(6), 419–431.

Edmondson, A. C. and R. Gulati (2021). Agility hacks how to create temporary teams that can bypass bureaucracy and get crucial work done quickly. *Harvard Business Review 99*(6), 46–49.

Ewens, M., R. Nanda, and M. Rhodes-Kropf (2018). Cost of experimentation and the evolution of venture capital. *Journal of Financial Economics 128*(3), 422–442.

Franzoni, C., P. Stephan, and R. Veugelers (2022). Funding risky research. *Entrepreneurship and Innovation Policy and the Economy 1*(1), 103–133.

Gallo, S. A., J. H. Sullivan, and S. R. Glisson (2016, 10). The influence of peer reviewer expertise on the evaluation of research funding applications. *PLOS ONE 11*(10), 1–18.

Graff Zivin, J. and E. Lyons (2021). The effects of prize structures on innovative performance. In *AEA Papers and Proceedings*, Volume 111, pp. 577–81.

Kerr, W. R., J. Lerner, and A. Schoar (2014). The consequences of entrepreneurial finance: Evidence from angel financings. *The Review of Financial Studies 27*(1), 20–55.

Krieger, J., D. Li, and D. Papanikolaou (2021, 03). Missing Novelty in Drug Development*. *The Review of Financial Studies*. hhab024.

Krieger, J. L. (2021). Trials and terminations learning from competitors rd failures. *Management Science 67*(9), 5525–5548.

Lane, J. N., M. Teplitskiy, G. Gray, H. Ranu, M. Menietti, E. C. Guinan, and K. R. Lakhani (2021). Conservatism gets funded? a field experiment on the role of negative information in novel project evaluation. *Management Science*.

Lee, C. J. (2015). Commensuration bias in peer review. *Philosophy of Science 82*(5), 1272–1283.

Lerner, J. and R. Nanda (2020). Venture capital's role in financing innovation: What we know and how much we still need to learn. *Journal of Economic Perspectives 34*(3), 237–61.

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

Malenko, A., R. Nanda, M. Rhodes-Kropf, and S. Sundaresan (2021). Investment committee voting and the financing of innovation. *Harvard Business School Entrepreneurial Management Working Paper* (21-131).

Morgan, P., D. G. Brown, S. Lennard, M. J. Anderton, J. C. Barrett, U. Eriksson, M. Fidock, B. Hamren, A. Johnson, R. E. March, et al. (2018). Impact of a five-dimensional framework on r&d productivity at astrazeneca. *Nature reviews Drug discovery 17*(3), 167–181.

Myers, K. (2020). The elasticity of science. *American Economic Journal: Applied Economics 12*(4), 103–34.

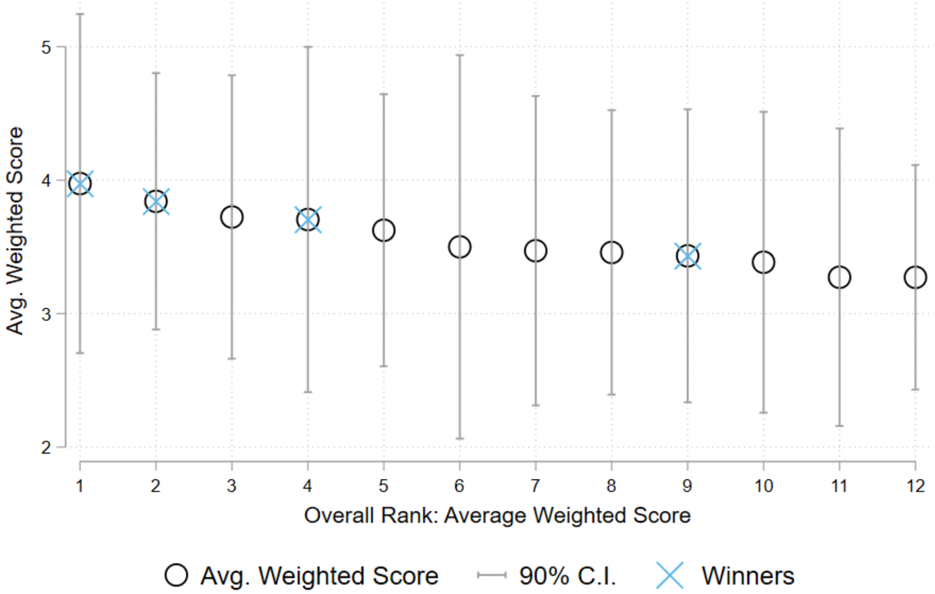
Nanda, R. and M. Rhodes-Kropf (2013). Investment cycles and startup innovation. *Journal of Financial Economics 110*(2), 403–418.

Nanda, R. and M. Rhodes-Kropf (2017). Financing risk and innovation. *Management Science 63*(4), 901–918.

Shih, H.-P., X. Zhang, and A. M. Aronov (2018). Drug discovery effectiveness from the standpoint of therapeutic mechanisms and indications. *Nature Reviews Drug Discovery 17*(1), 19–33.

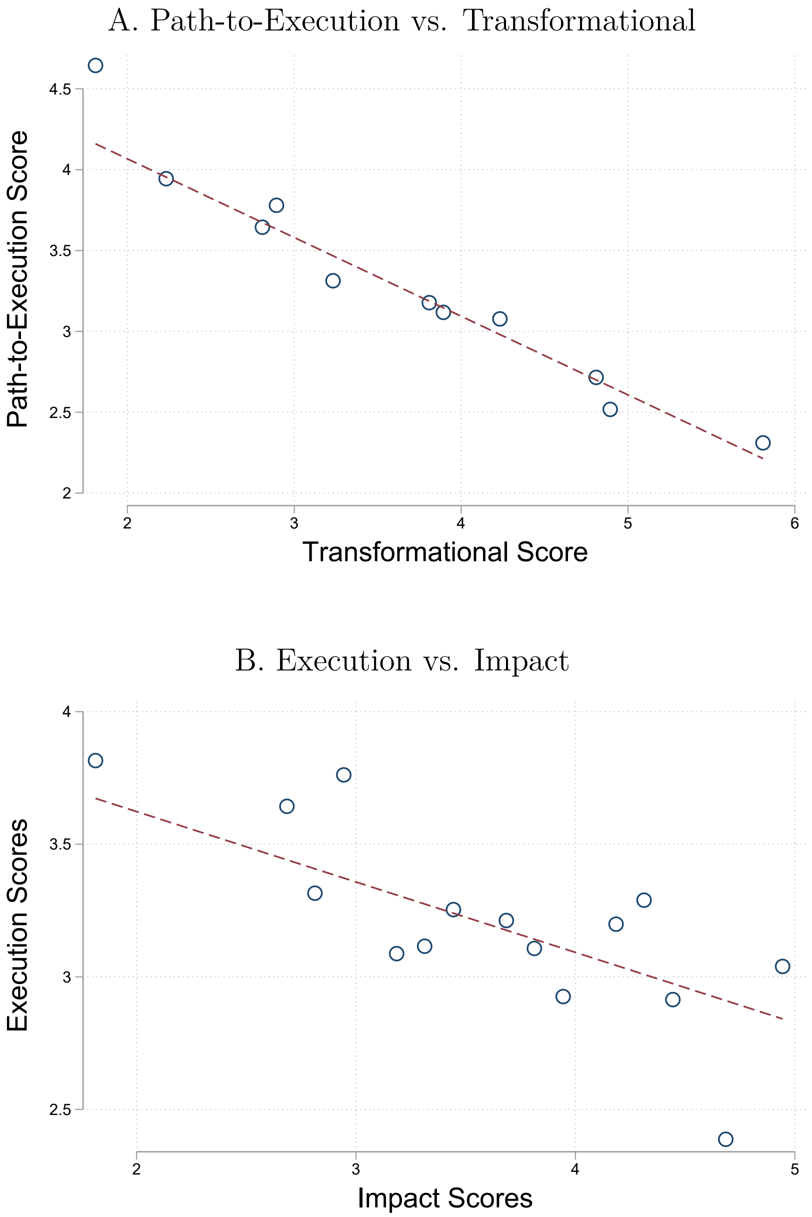
Tables & Figures

Figure 1: Weighted Average Scores, by Overall Rank (Survey #1)



Notes: Figure 1 graphs the Pitch Day participants' average weighted scores ordered by rank (best to worst average scores), and with the 90% confidence interval (grey bars). The winning projects chosen by the selection panel are marked by Xs. The weighted scores give extra weight to certain attribute scores (Transformational Potential: 3x, Feasibility / Path-to-Execution: 2x, Team: 2x).

Figure 2: Tradeoff Between Impact Potential and Execution Risk (Survey #1)

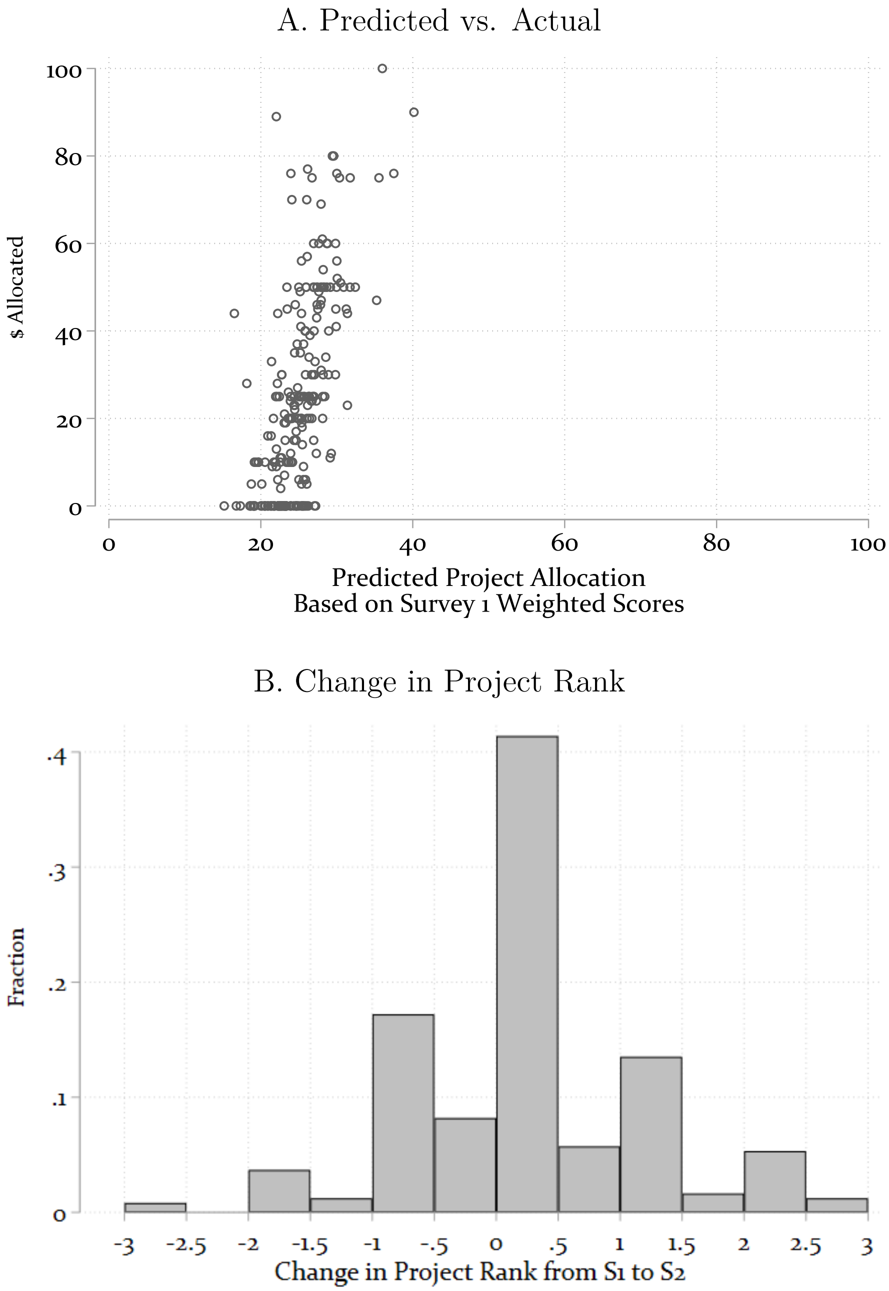


A. Path-to-Execution vs. Transformational

B. Execution vs. Impact

Notes: Figure 2 shows the correlation between project attributes, within Pitch Day participant's project evaluations (Survey #1). Panel A compares the Path-to-Execution score to the Transformational Potential score. Panel B adds the the Path-to-Execution score and Timescale score together into the "Execution Scores" and the Transformational Potential and Breadth scores into the "Impact Scores." Both binscatter plots show display equally sized quantiles of the x-axis variable and include all 244 participant-evaluations from Survey #1. The binscatters additionally control for tercile of the project participant's average weighted score, such that the correlation displayed can be interpreted as the within-scoring tercile relationship between those attribute groups.

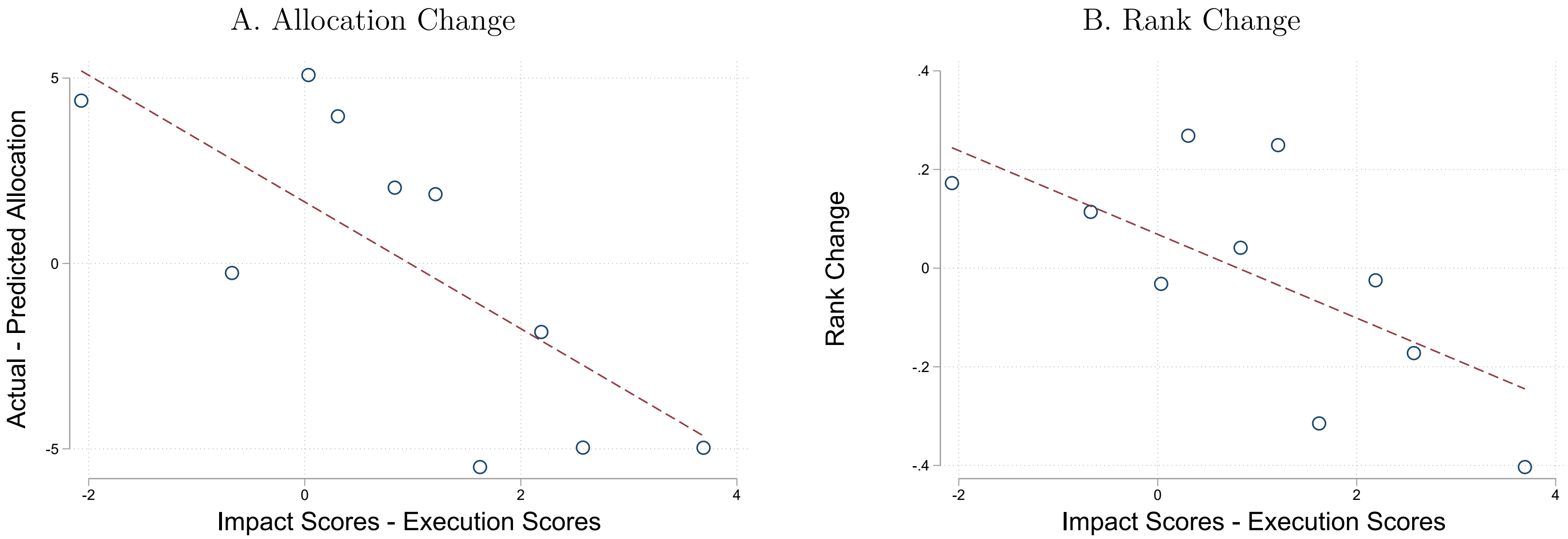
Figure 3: Portfolio Evaluation (Survey #2), Predicted vs. Actual Scores


A. Predicted vs. Actual

B. Change in Project Rank

Notes: Figure 3 shows the results of the portfolio allocation decisions (Survey #2). Panel A graphs the distribution of actual allocations by predicted project allocations, where $Predicted\hat{A}llocation = \frac{WScore_{i,j}}{\sum W.Score_i} \times 100$. Panel B graphs the distribution of all project rank changes from Survey #1 to Survey #2 for the 61 participants (244 participant-projects) completed in Survey #2.

Figure 4: Portfolio Evaluation (Survey #2): Impact vs. Execution Penalty



Notes: Figure 4 shows binscatter plots of the change from predicted allocation (Panel A) and rank order (Panel B) to actual project allocation and rank order in Survey #2. Both panels display results by quantile of participant-project attribute differences between Impact scores (Transformational Potential and Breadth added together) and Execution scores (Path-to-Execution and Timescale). The y-axis in Panel A is the difference between the evaluator-project predicted dollar allocation (based on Survey #1 average weighted score) and that participant's actual dollar allocation in Survey #2. Predicted allocation is calculated as $Predicted\hat{A}llocation = \frac{WScore_{i,j}}{\sum W.Score_i} \times 100$. In Panel B, the y-axis shows change in relative rank between the focal project and the others ranked by that same participant in the same session—e.g., if the participant's Survey #1 average weighted scores suggested a project would be ranked first among the four projects in a session, but their Survey #2 allocations had that project ranked second, then the rank change would equal -1; if the project went from being tied for second to third ranked, then the change would be -0.5. Both binscatters control for participant-project expertise, and implied rank (Survey #1).

Online Supplement

A Genesis Labs Evaluation and Selection

For the particular request for proposals that we studied, the program's focus was on "therapeutic-enabling technologies" that would otherwise be outside the scope of an existing department. Teams of scientists within the company could submit a short description of their ideas. In particular, the solicitation encouraged proposals that would not fall into the normal project funding pathways in the following areas:

- *New breakthrough therapeutic technologies*: enabling drugging upreecedented targets; accelerating drug hunting, development, and trials.
- *Reimagined technologies*: exploiting existing technologies to identify solutions for previously intractible therapeutic concepts; redeploying existing technologies to remove bottlenecks in drug research and development
- *Data, digital, and automation*: New methods using existing data to identify targets, conduct drug discovery or developent; in-silico strategies (computer simulations) to streamline eisting drug discovery projects; digital tools to improve patient outcomes or enhance existing assets.
- *Patient centric medicine*: New ways to use targeted delivery approaches to improve patient compliance
- *White space*: What else are we not doing today, that we should be doing? Disruptive out of the box therapeutic concepts that fall outside of our existing strategy.

A.1 Genesis Labs Evaluation Process

After an initial triage process, in which the core program team[10] removed non-viable applications that did not fit in the program, 55 teams were invited to further develop

[10] Three of this study's authors (Hunt, Reynolds and Tarsa) were members of the core program team

their project proposals and submit them to a peer review process. 146 company scientists participated in the peer review process of 51 final proposals.[11] Reviewers were assigned based on their scientific discipline relevance to the project. Between four and seven reviewers read and scored each proposal, resulting in 299 reviews. Reviewers submitted a score (between 1–5) for five categories:[12]

- *Transformational Potential (3x weight)*: The proposal should be creative, non-standard and have the potential to open up novel research directions.
- *Breadth of Applicability*: Proposals should have high value proposition to the organization and offer new solutions to previously unmet medical need
- *Timescale to First Prototype*: Path to internalization should be within reason, aligned to the 18 month funding period allocated to winning teams
- *Feasibility / Path to Execution (2x weight)*: Even though unproven, the concept should be credible (e.g., robust flow chart).
- *Team (2x weight)*: Expertise, network and quality of the team

The core review team, consisting of roughly 25 leading internal scientists and the program leaders aggregated the peer review data and convened over two sessions in June 2020. Guided by the review scores, the committee discussed the relative merits of the projects and selected the projects for the next stage.[13] This competitive process ensured that the final Pitch Day projects were already at a high quality baseline, such that a set of experienced company

[11] A number of teams merged their proposals between the initial submission and peer review.

[12] In both the peer review phase and the final Pitch Day scoring, evaluators were told that some categories would be weighted double (Path-to-Execution, Team) or triple (Transformational Potential) in the evaluation's overall score.

[13] Conversation about projects was mostly an open, unstructured forum, and varied across discussion of project scientific qualities and fit with the program's goals (i.e., would the project be funded within more traditional project pathways?). The peer review scores were shared in those discussion, but ultimate decisions were not formally beholden to consensus scores.

scientists believed that each shortlisted project was worthy of spotlight in front of company leaders, scientists from the firm's multiple global research offices and an external academic expert. The peer review process would have filtered out projects with no relevance to the Genesis Labs program goals, without any scientific merit, or lacking any feasible path to success.

The core team selected 12 proposals to pitch their ideas to a panel of eight company leaders and one external academic expert.

Attendees participated in project evaluation on their personal computer or smart phone. The organizers invited attendees in the pre-event emails and with QR codes and links shared at the start of each session. The project evaluations were collected in a series of "live" digital surveys that participants filled out throughout the pitch sessions at specified intervals.[14] In the introductory welcome remarks to each session, the designated master of ceremonies invited attendees to participate in the surveys by clicking on the link from the pre-event emails or using a QR code displayed on the screen. Upon entering the (Qualtrics) survey interface, participants saw brief instructions—including the project score categories and the formula used to calculate weighted scores (see Figure A.1, Panel B)– and asked to fill out some basic information: years at the company, division within the company, years in the industry, and field of highest degree earned.

The smaller panel of firm leaders and one outside academic scientists convened after the sessions to select four winning projects to fund. The official selection panel had backgrounds quite similar to the audience participants (i.e., PhD scientists, MDs, and biopharma industry veterans). The selection panel did not have access to the audience scores.

[14] Organizers mentioned that the evaluation surveys were part of a collaboration with Harvard researchers to study project evaluation.

A.2 Pitch Day Scoring Process

At the end of the presentation (before the Q&A), the audience evaluators were instructed to fill out their independent evaluations.[15] The independent evaluations first asked the evaluator to rate their own level of expertise in evaluating the project. Next, participants score the project on a five-point Likert scale (from "1: Poor" to "5: Excellent") across the same five dimensions as the peer review scores: Transformational Potential, Path-to-Execution, Breadth of Applicability, Team, and Timescale-to-Prototype.

At the end of the last pitch in a session, participants were automatically redirected to the Portfolio Allocation scoring interface (Survey #2). The instructions read as follows: "Which projects should [the program] support? Assume you have 100 "dollars" to invest. Please allocate those 100 dollars across the session's projects." The evaluators then could move a set of sliders (one for each project) between 0–100, and the total allocations had to add up to 100.

To remind the participants of their independent evaluations, a table above the sliders reported the participant's weighted scores for each project based upon his or her Survey #1 category scores.[16] This reminder and its prominent placement are critical, because they ensure that participants are fully informed about their own independent project weighted scores and their differences across projects.

[15] Using timestamps from survey clicks/entries, we discard participant data when the participant did not abide by this timeline.

[16] Half of the participants were randomized into an additional treatment arm, in which the table also showed the "crowd" average weighted score (based on their peers' aggregated Survey #1 responses). We analyze the results of this additional experiment in another paper. In this paper, we control for that peer treatment status in all analyses of participants' portfolio allocation choices.

B Descriptive Statistics

Table A.1, Panel A reports the composition of the participants.[17] In response to the question "In which field did you earn your highest degree?", 35% responded biology or biochemistry, 32% chemistry, 11% medicine or pharmacy, and 22% in data science or business. The composition was very similar (+/- 5%) across sessions. Participants' tenure at the company ranged from 0 to 30 years, with a mean of 11 years at the company.

Table A.1, Panel B compares participants who only completed the category scoring portion of the session's scoring (Survey #1) with the group have also completed the dollar allocation portion (Survey #2). The composition of researcher fields appears quite similar, with the exceptions of relatively fewer medical doctors and pharmacists and relatively more biologists in the group that did not complete all steps for the resource allocation analysis. That group also has slightly greater self-reported project expertise.

Participants were required to rate their own level of expertise with respect to evaluating each separate project on a five point Likert scale ("1: Totally unfamiliar, "2: Mostly unfamiliar," "3: Generally familiar, but never worked in this area myself," "4: Somewhat related to my area of work/expertise," "5: Directly related to my area of work/expertise"). We note that this self-reported measure of expertise varied within participants at the project level. The distribution of expertise scores was rather symmetric (see Figure A.7, Panel B). The mean response was a 3.05 out of 5, with a mode of 3 (35.67% responses were below 3 and 33.3% were above 3). Throughout the analysis below, we refer to self-reported "experts" as participant-project pairs with an expertise score of 4 or 5.

[17] Some participants repeated across pitch sessions. Since our surveys did not require participants to identify themselves by name, we cannot be sure exactly how many were repeat session participants. Using survey demographics responses we estimate that we had 105 unique participants, 25 of whom participated in more than one pitch session.

C Attribute Scoring Patterns

In Figure A.9, we drill further into which qualities drive differences in attribute inconsistency by looking at specific pairs of attributes. First, we account for level differences between categories and participants by normalizing all attribute scores based on attribute and evaluator fixed effects.[18] Next we generate all the pairwise differences of within-evaluation attribute scores, such that every participant-project-evaluation is represented by 10 pairs (attribute differences). Panel A shows the distribution of these normalized attribute pair score differences.

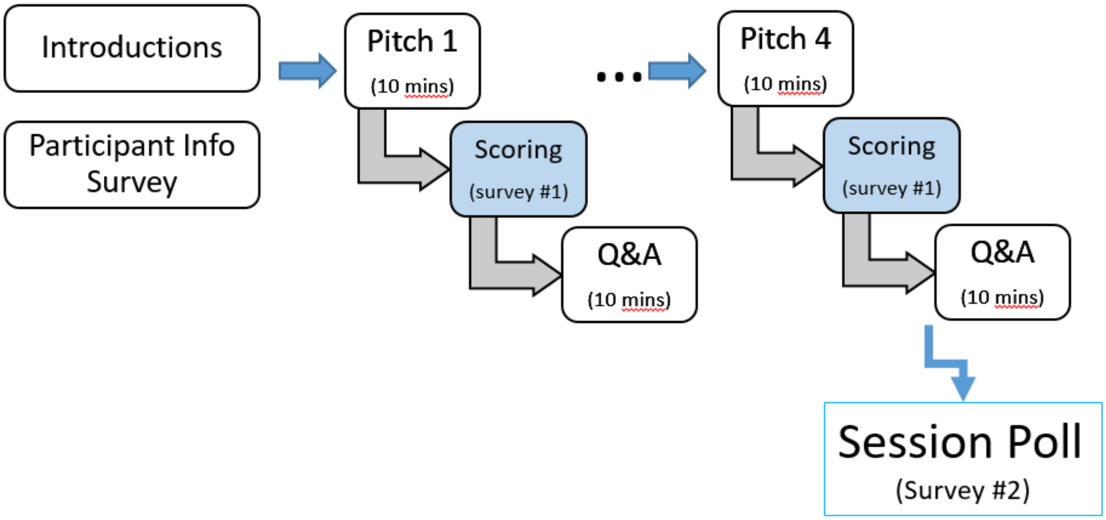
Looking at the pairwise differences reveals that attribute inconsistency is much more likely to be driven by difference in project practicality and potential impact measures. The average gap between residualized category scores is 60% larger for the pair with the largest average gap (Transformational–Path-to-Execution) than it is for the group with the lowest (Path-to-Execution–Time). 226 (45%) of all evaluations have the greatest (residualized) within-evaluation gap between their Transformational score and another category. Of the Transformational score pairs, 33% had the largest gap with the Timescore, 31% with Path-to-Execution, 24% with Team and 12% with Breadth. Outside of the transformational score, the category next most likely to be involved in the biggest within evaluation pairwise gap was path to execution (41%), followed by Time (39%), Team (38%), and Breadth (37%). Figure A.9, Panel B shows the pairwise difference distributions for the two pairs with the highest and lowest average inconsistency. Unsurprisingly, the attributes pairs that evaluate probability of success (Path-to-Execution and Timescale) and those that track potential impact (Transformational Potential and Breadth) show the most positive correlations, while

[18] We can only speculate about the reasons for different distributions in raw attribute scores (see Panel A of Figure A.7). For example, the "Transformational Potential" scores probably skewed higher because of both selection (i.e., the peer review round emphasized this category) and salience effects (the participant instructions highlighted the extra weight put on this category. Even in an anonymous scoring process, one can easily imagine that professional colleagues would feel more comfortable assigning low scores to project path to execution and Timescale than to Team quality.

the cross between those two groups shows the greatest pairwise dispersion.

Figure A.1: Experiment Flow, Instructions, Scoring Interface

A. Sequence of Events



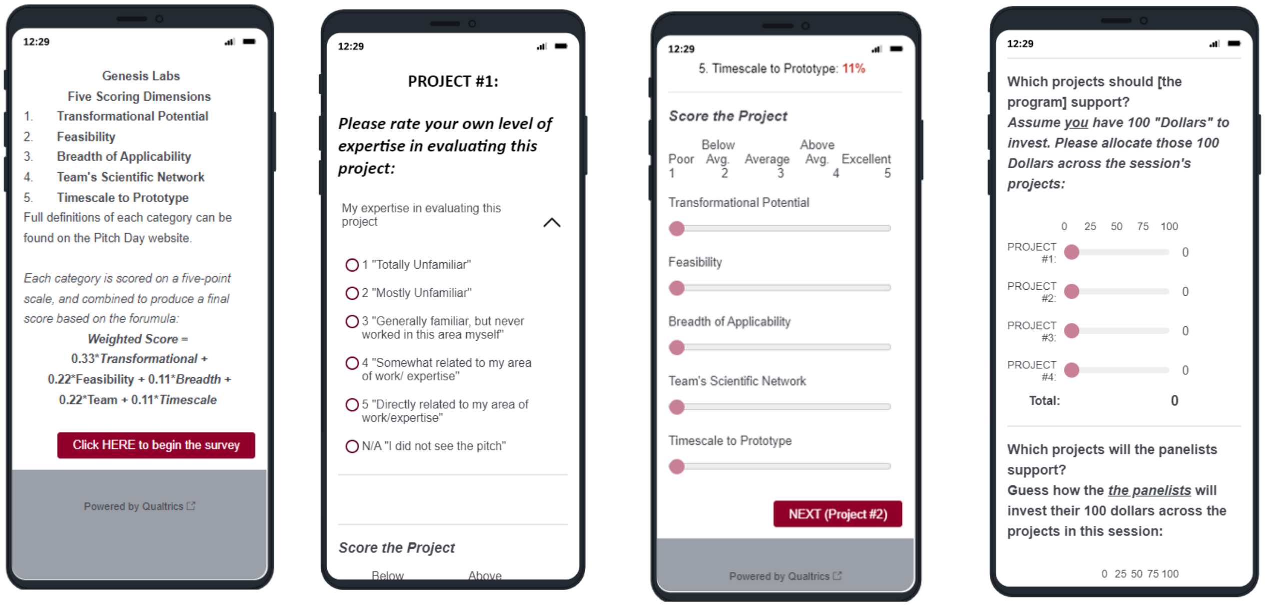
B. Scoring Interface (for smartphones)

Scoring

Expertise

Indep. Evaluation (S1)

Portf. Allocation (S2)



12:29

Genesis Labs

Five Scoring Dimensions

1. Transformational Potential
2. Feasibility
3. Breadth of Applicability
4. Team's Scientific Network
5. Timescale to Prototype

Full definitions of each category can be found on the Pitch Day website.

Each category is scored on a five-point scale, and combined to produce a final score based on the forumula:

***Weighted Score** =*

0.33*Transformational +

0.22*Feasibility + 0.11*Breadth +

0.22*Team + 0.11*Timescale

Click HERE to begin the survey

Powered by Qualtrics

12:29

PROJECT #1:

Please rate your own level of expertise in evaluating this project:

My expertise in evaluating this project

○ 1 "Totally Unfamiliar"

○ 2 "Mostly Unfamiliar"

○ 3 "Generally familiar, but never worked in this area myself"

○ 4 "Somewhat related to my area of work/ expertise"

○ 5 "Directly related to my area of work/expertise"

○ N/A "I did not see the pitch"

Score the Project

Below Above

12:29

5. Timescale to Prototype: 11%

Score the Project

Poor 1	Below Avg. 2	Average 3	Above Avg. 4	Excellent 5

Transformational Potential

Feasibility

Breadth of Applicability

Team's Scientific Network

Timescale to Prototype

NEXT (Project #2)

Powered by Qualtrics

12:29

Which projects should [the program] support?

Assume <u>you</u> have 100 "Dollars" to invest. Please allocate those 100 Dollars across the session's projects:

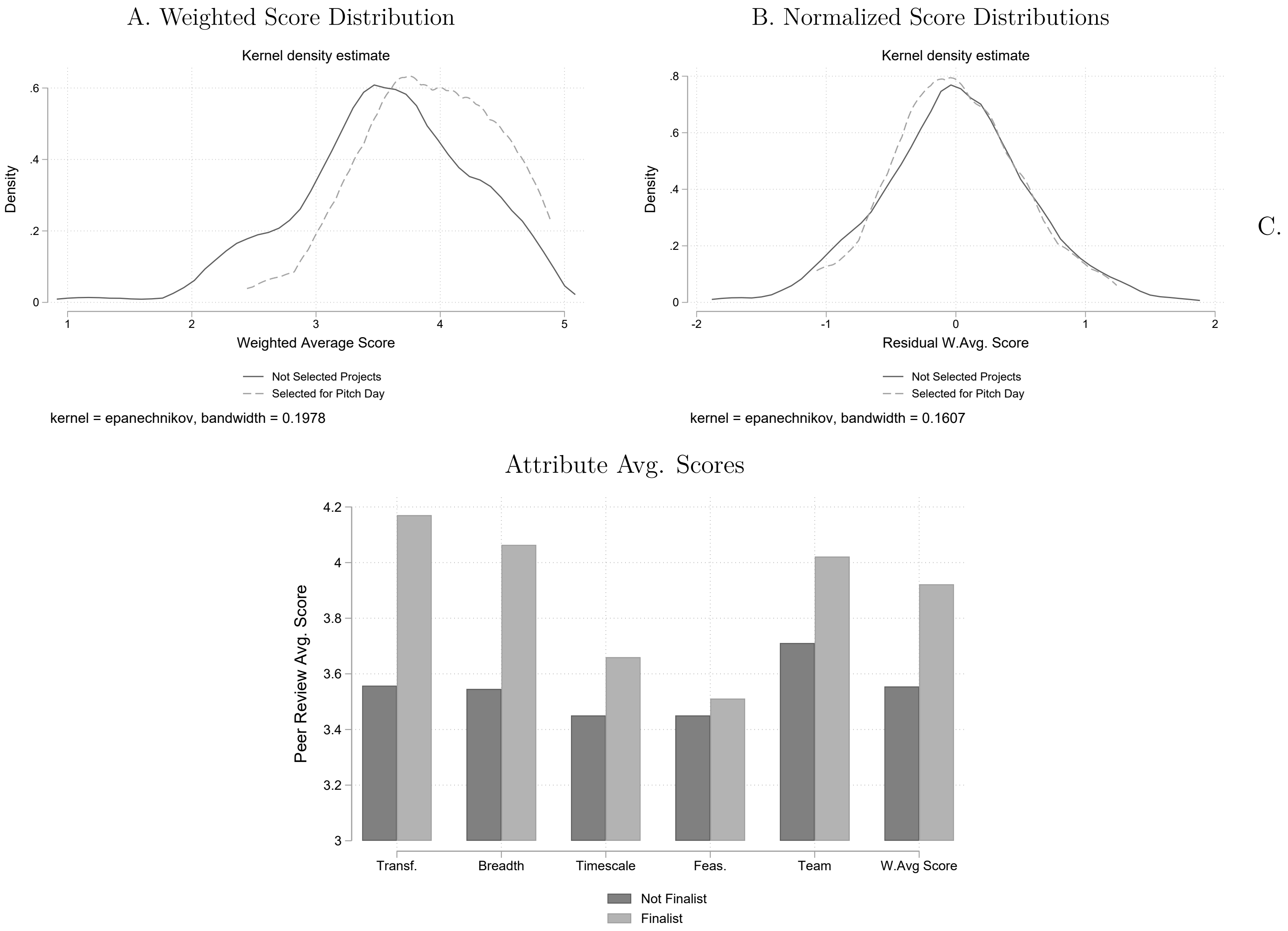
	0 25 50 75 100	
PROJECT #1:		0
PROJECT #2:		0
PROJECT #3:		0
PROJECT #4:		0
Total:		**0**

Which projects will the panelists support?

Guess how the *the panelists* will invest their 100 dollars across the projects in this session:

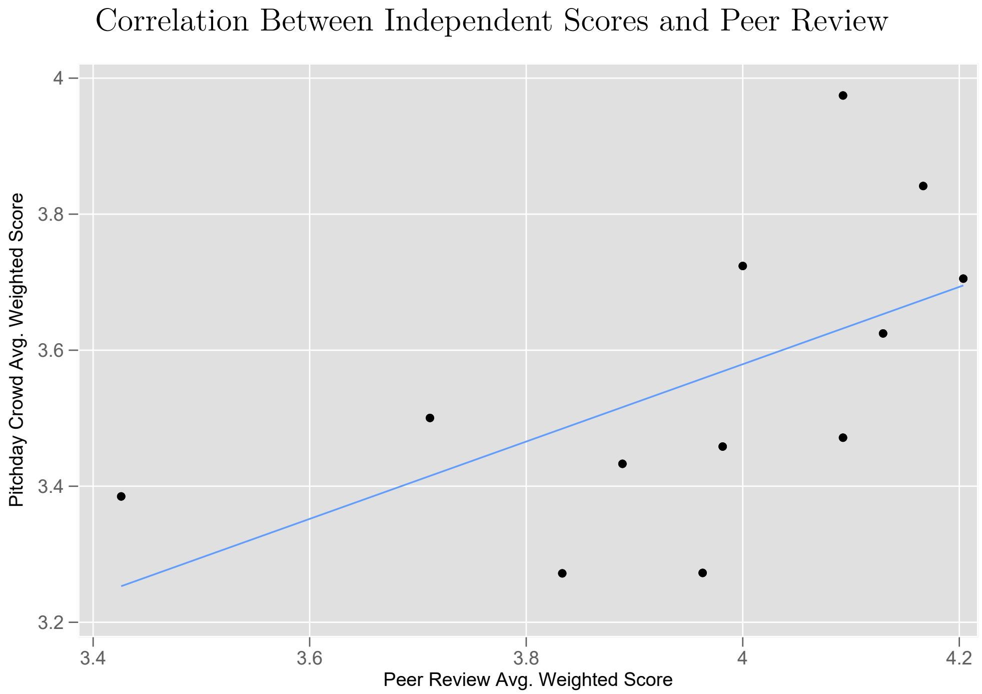
0 25 50 75 100

Figure A.2: Peer Review Scoring



Notes: Figure A.2 describes the results of the peer review evaluation of 51 semi-finalist Genesis labs projects. Each panel shows the results for the 39 non-finalists vs. the 12 selected finalists. Panel A displays weighted average score distributions (by participant-project). Panel B shows the same scores, but normalized by each projects average score. Panel C shows the average peer review scores by attribute and finalist status. Panel D shows the standard deviation (across peer reviewers and projects) for each attribute.

Figure A.3: Semifinal Peer Review vs. Finalist Pitchday Audience Scoring



Notes: Figure A.3 shows the correlations between each project's average weighted score from the earlier peer review round and its average weighted score based on the Pitch Day audience evaluations based on the sample of 141 participants and 503 participant-project evaluations. The weighted scores give extra weight to certain attribute scores (Transformational Potential: 3x, Feasibility / Path-to-Execution: 2x, Team: 2x).

Table A.1: Summary Statistics

Evaluators	Mean
Company Tenure (years)	11.13
10+ Years in Industry	0.72
Field: Biology/Biochem.	0.35
Field: Chemistry	0.32
Field: Medicine/Pharmacy	0.11
Field: Business/Data Science	0.04
Field: Other	0.18
Program Reviewer (Ever)	0.26
Min. Project Expertise	2.17
Avg. Project Expertise	3.11
Max. Project Expertise	4.01

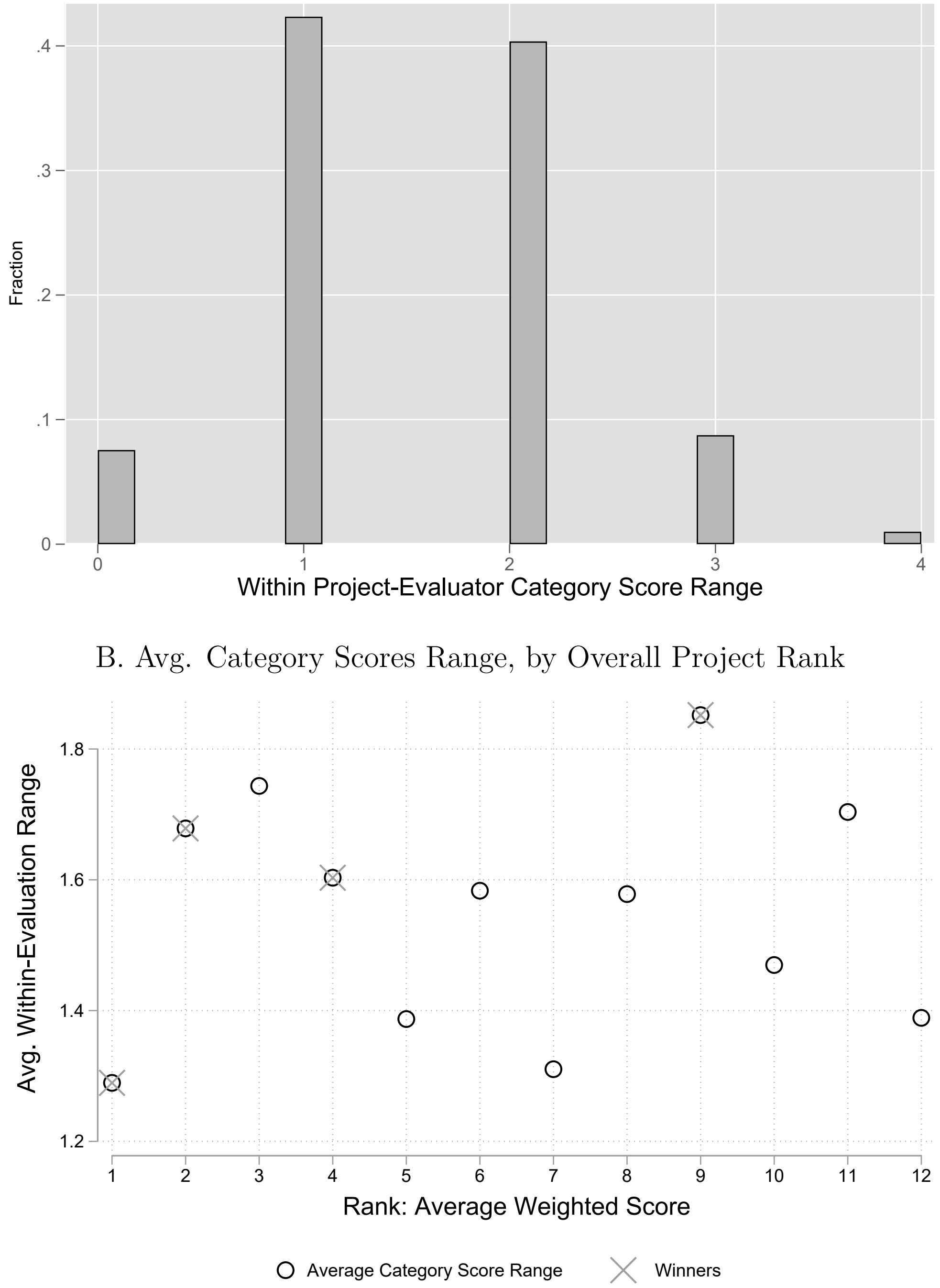
Notes: Table A.1 presents summary statistics from the sample of 141 participants in Survey #1.

Table A.2: Comparing Survey Samples

	Survey #1 Only	Survey #1 and Survey #2	t-test p-value
	N=80	N=61	
Company Tenure (years)	10.14	12.43	0.11
10+ Years in Industry	0.71	0.72	0.91
Field: Biology/Biochem.	0.30	0.43	0.13
Field: Chemistry	0.36	0.26	0.20
Field: Medicine/Pharmacy	0.16	0.05	0.03**
Field: Business/Data Science	0.01	0.07	0.13
Field: Other	0.16	0.20	0.61
Genesis Reviewer (Ever)	0.24	0.30	0.45
Min. Project Expertise	2.25	2.07	0.27
Avg. Project Expertise	3.24	2.93	0.05**
Max. Project Expertise	4.16	3.80	0.06**

Notes: Table A.2 compares the composition of the participants who only completed the category scores in Survey #1 to those who fully completed both surveys for their pitch session (i.e., scored all four projects by category in Survey #1, then completed the resource allocation questions in Survey #2). The last column reports the p-value from a two sample t-test of the two groups (** $p < 0.05$, * $p < 0.1$). The comparisons show that the Survey #1 only group included significantly more MDs and PharmD, and slightly less Biologists and Biochemists. Self reported expertise was also slightly higher in the Survey #1 only group.

Figure A.4: Range Across Evaluation Categories (Within Evaluation)

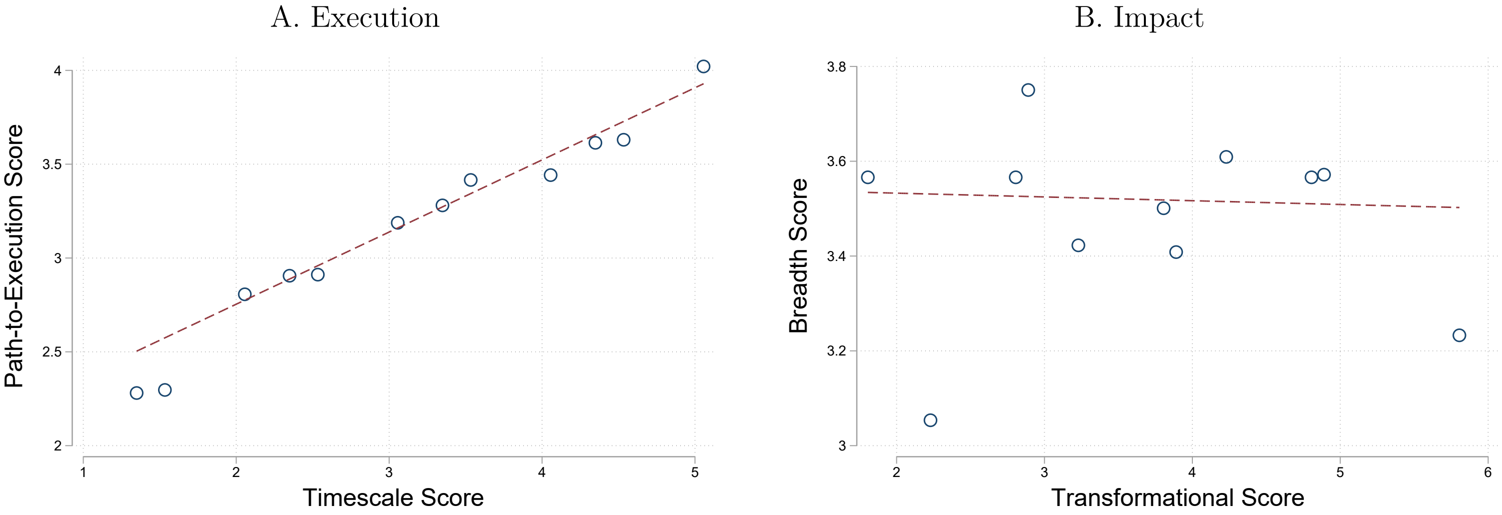


A. Category Scores Range (Within Evaluation)

B. Avg. Category Scores Range, by Overall Project Rank

Notes: Figure A.4 Panel A shows the distribution of attribute range (max - min) across all participant-project evaluations. All attributes were scored on a 1–5 Likert scale, so the maximum range is capped at 4. Panel B displays the average attribute range by project rank. The winning projects chosen by the selection panel are marked by Xs.

Figure A.5: Attribute Tradeoffs, Within Execution and Impact Groups

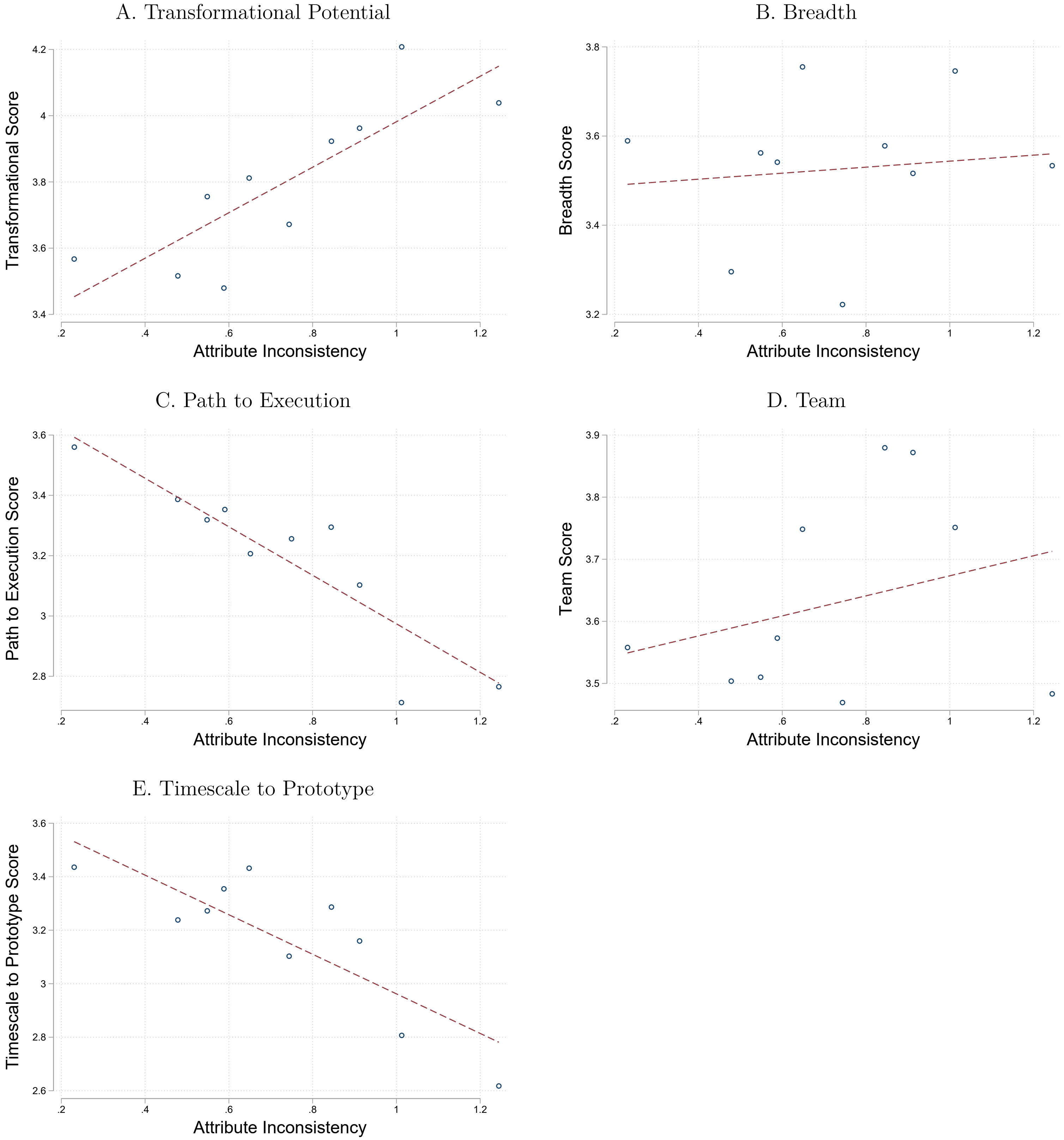


A. Execution

B. Impact

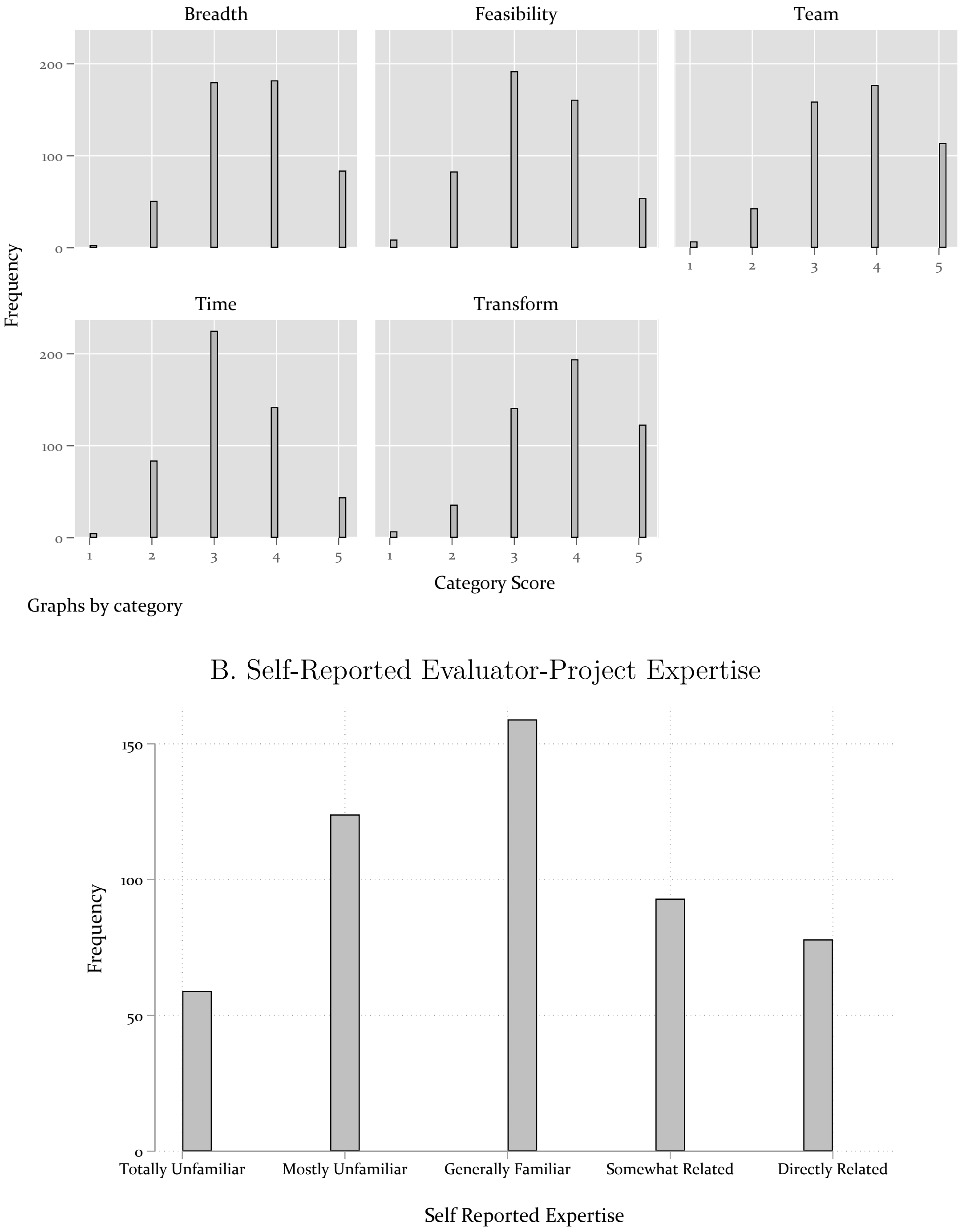
Notes: Figure A.5 shows the correlation between project attributes, within Pitch Day participant's project evaluations (Survey #1). Panel A compares the two "Execution" scores (Path-to-Execution and Timescale score), while Panel B compares the two "Impact" scores (Transformational Potential and breadth). Both binscatter plots show display equally sized quantiles of the x-axis variable and include all 244 participant-evaluations from Survey #1. The binscatters additionally control for tercile of the project participant's average weighted score, such that the correlation displayed can be interpreted as the within-scoring tercile relationship between those attribute groups.

Figure A.6: Conflicted Evaluations and Attribute Scores



Notes: Figure A.6 show the correlation between an evaluation's overall attribute imbalance—the standard deviation across a given participant-project's five category scores—and each one of those scores. Each panel displays the correlations as binscatters by decile of attribute imbalance, after adjusting for project fixed effects and evaluator-project characteristics (expertise and "home team" bias).

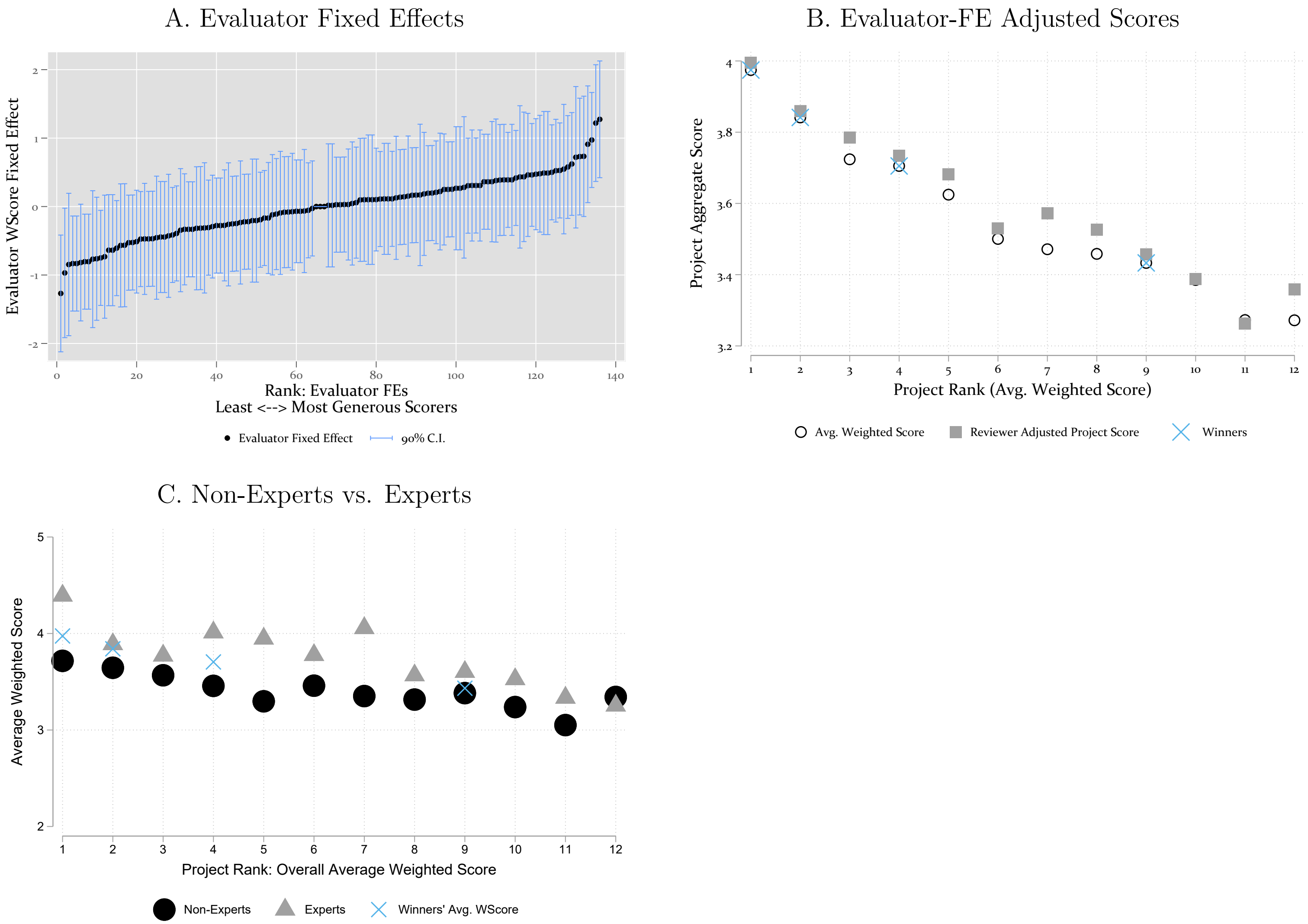
Figure A.7: Distribution of Project–Category Scores (Survey #1)


A. Independent Evaluation Category Scores

B. Self-Reported Evaluator-Project Expertise

Notes: Figure A.7 Panel A displays the project category score distributions from Survey #1 (independent scoring). The sample includes 141 session participants, 504 participant-project evaluations, and 2500 participant-project-category scores. The mean score across categories is 3.53.

Figure A.8: Independent Scoring (Survey #1) Outcomes: Evaluator and Expertise Effects



A. Evaluator Fixed Effects

B. Evaluator-FE Adjusted Scores

C. Non-Experts vs. Experts

Notes: Figure A.8, Panel A graphs the evaluator scoring fixed effects. Evaluator fixed effects are calculated separately for each session's project evaluations. Each point represents an evaluator's average category score generosity relative to the median evaluator. The blue bars represent 90% confidence intervals. Panel B shows how adjusting for those evaluator fixed effects slightly changes projects' average weighted scores and ranks. Panel C shows how the average weighted scores differ for self-identified experts and non-experts.

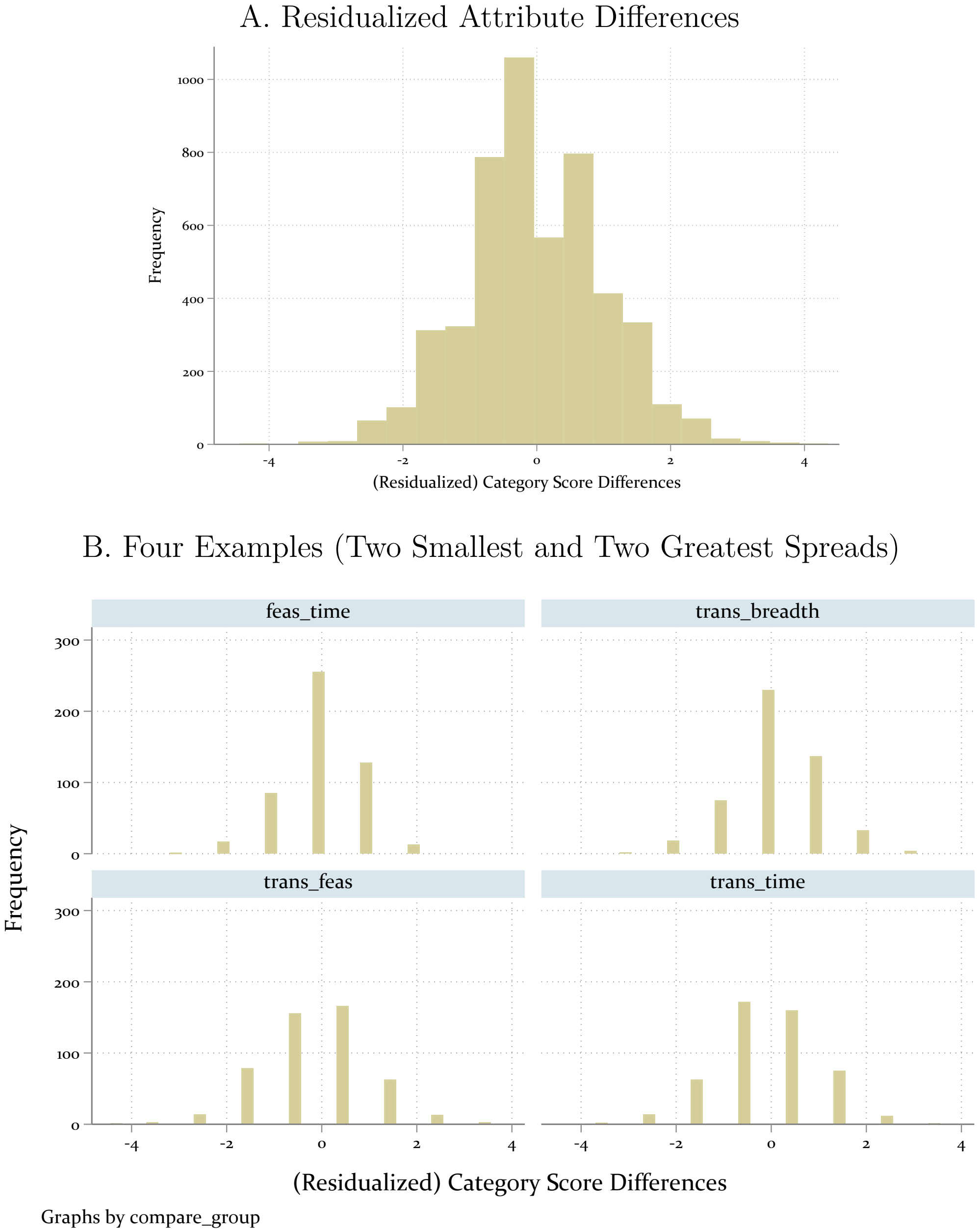
Table A.3: Independent Category Scoring Regressions

VARIABLES	(1) Value	(2) Value	(3) Value	(4) Value	(5) Value
group(expertise) = 2					-0.276**
					(0.107)
group(expertise) = 3					-0.0959
					(0.103)
group(expertise) = 4					0.171
					(0.119)
group(expertise) = 5					0.357***
					(0.136)
Same Org (Project-Evaluator)	0.162**	0.0983	0.0537	0.166*	0.0610
	(0.0807)	(0.0792)	(0.0855)	(0.0860)	(0.0814)
Expert	0.308***	0.424***	0.388***	0.286***	
	(0.0686)	(0.0691)	(0.0663)	(0.0661)	
Tenure (years)				-0.00907**	
				(0.00357)	
Observations	2,486	2,486	2,486	2,486	2,486
R-squared	0.074	0.278	0.313	0.120	0.320
Category FE	YES	YES	YES	YES	YES
Participant FE		YES	YES		YES
Project FE			YES	YES	YES

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

Notes: Appendix Table A.3 reports the ordinary least squares regressions of attribute value at the evaluator-project level. The independent variables are various evaluator and evaluator-project specific characteristics. Columns 1–4 include a indicator variable for whether the evaluator was a self-reported expert (4/5 or 5/5) for the given project, and whether the evaluator came from the same R&D division as at least one of the project team members. Column 4 adds a running variable for evaluator tenure (in years) at the company, and Column 5 reports separate coefficient for each level of expertise (1 out of 5 is the omitted category). All models include attribute (category) fixed effects, while only some include participant fixed effects, and project fixed effects.

Figure A.9: Within-Evaluation Pairwise Differences

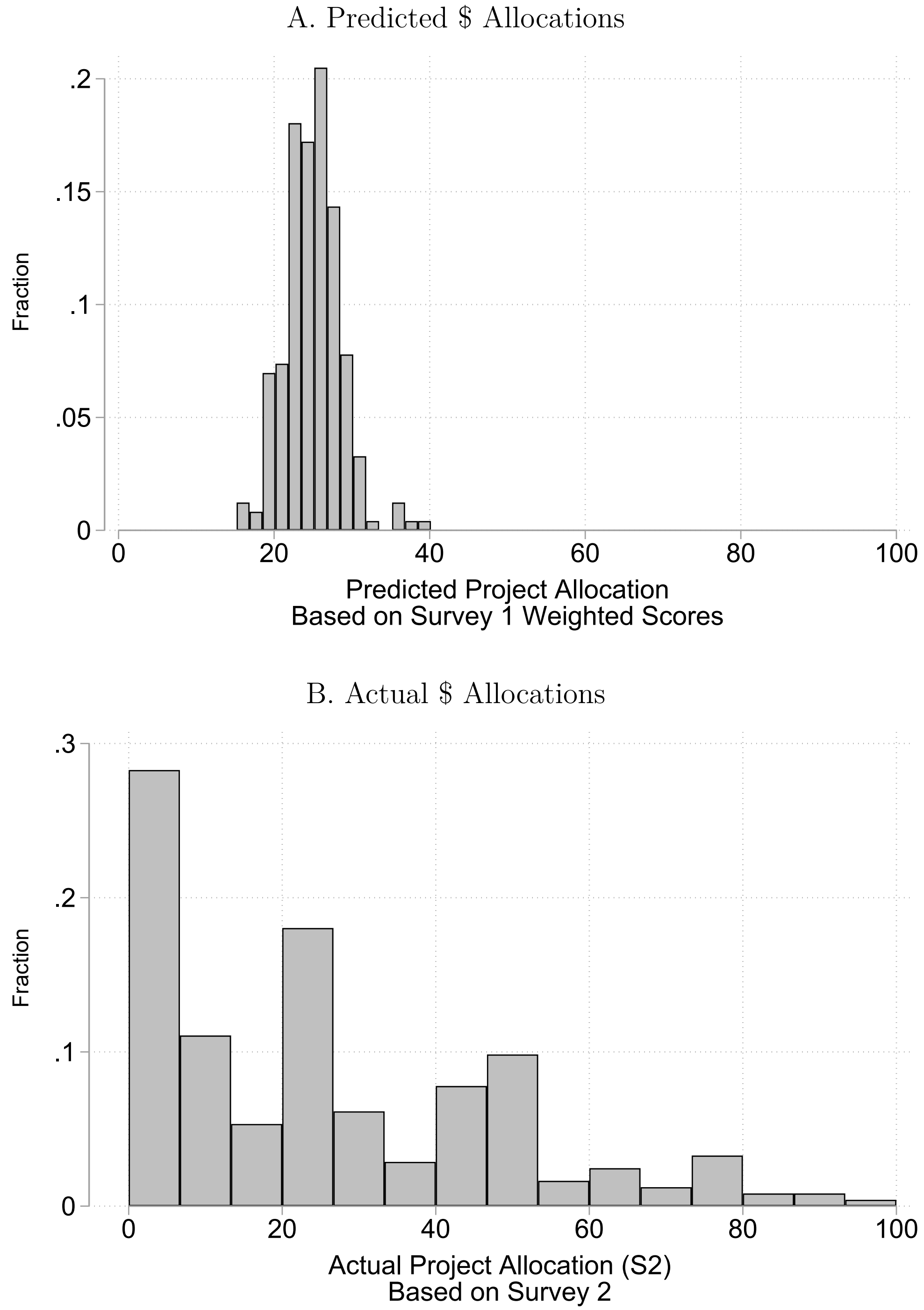


A. Residualized Attribute Differences

B. Four Examples (Two Smallest and Two Greatest Spreads)

Notes: Figure A.9 shows the histograms category score differences within participant-project-evaluations. The score are first normalized to account for category and evaluator fixed effects. Next, we calculate all pairwise differences between the (residualized) category scores within a participant-evaluation. For example, if the participant's residualized project scores were 4 for Transformational and 1 for Team, then their pairwise category difference would be 3 for Transformational–Team. Panel A reports the distribution of all pairwise within participant-project category score comparisons. To further illustrate the variation in these pairwise comparisons, Panel B shows the histogram of residualized category differences to the two groups with the least average differences between their within project-evaluator category scores (feasibility-time and transformational-breadth), and the two least synchronous pairs (transformational-feasibility and transformational-time).

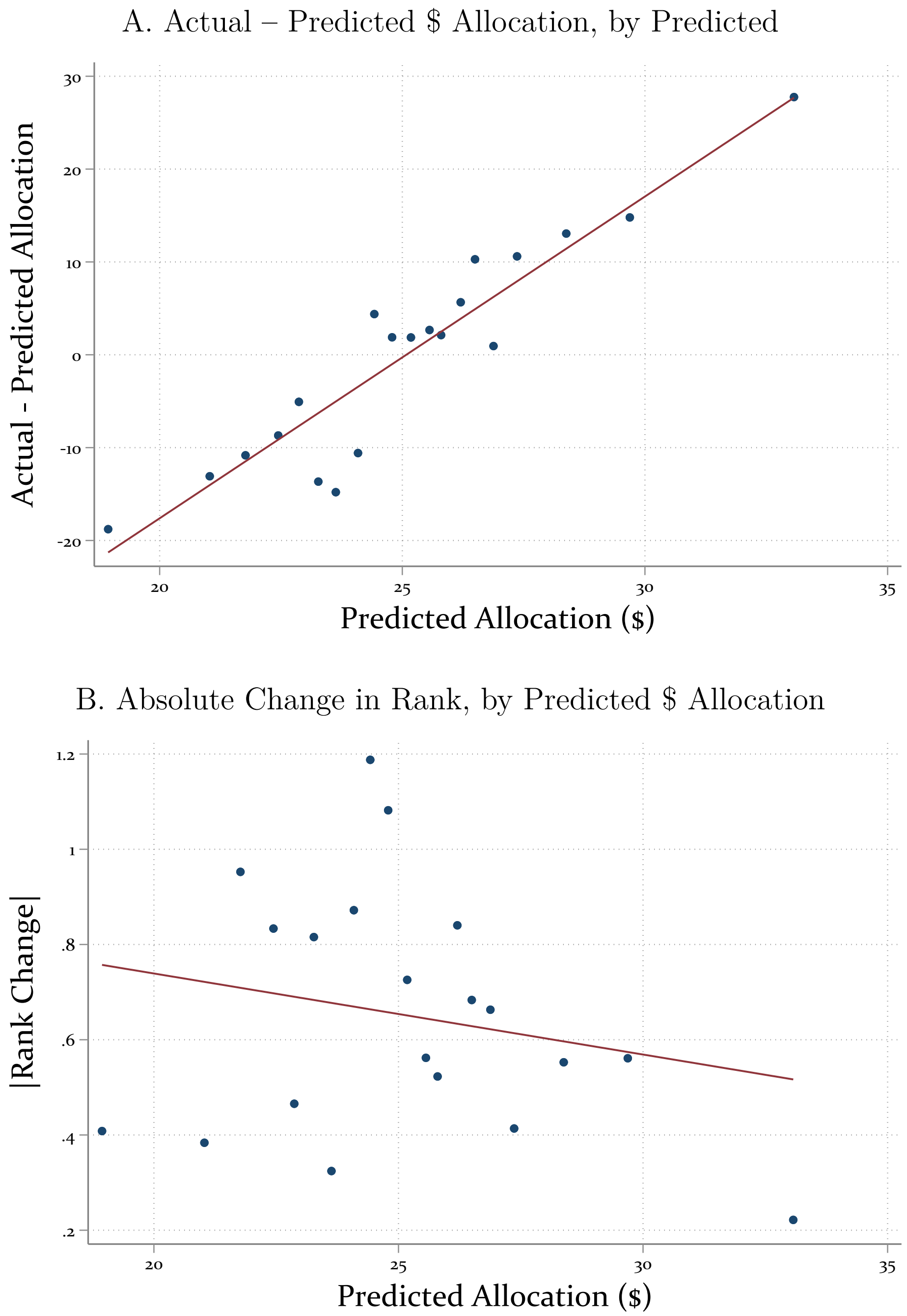
Figure A.10: Portfolio Evaluation (Survey #2) Predicted vs. Actual Allocation Distribution


A. Predicted $ Allocations

B. Actual $ Allocations

Notes: Figure A.10 shows the results of the portfolio allocation decisions (Survey #2). Panel A graphs the distribution of predicted project allocations, where $Predicted\hat{A}llocation = \frac{WScore_{i,j}}{\sum W.Score_i} \times 100$. Panel B graphs the distribution of all project allocation choices for the 61 participants (244 participant-projects) completed in Survey #2.

Figure A.11: Actual vs. Predicted Allocations and Ranks (S1 vs. S2), Binscatters



A. Actual – Predicted $ Allocation, by Predicted

B. Absolute Change in Rank, by Predicted $ Allocation

Notes: Figure A.11 shows binscatter plots by 20 quantiles of evaluator-project predicted allocations, based on the relative weighted scores of that evaluator's Survey #1 unconstrained scoring. The Y-axis in panel A is the difference between their actual allocations (Survey #2) and the predicted allocation. In Panel B, the outcome is the absolute value of that evaluator's rank order changes for the given project, among the set of four in the pitch session. Both plots adjust for participant-project characteristics (expertise and "home team" bias), as well as their evaluation's overall attribute inconsistency (standard deviation across the five attributes), and peer score treatment status.

Table A.4: Portfolio Allocation, by Predicted Values and Evaluation Characteristics

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
VARIABLES	Allocation	Allocation	Allocation	Allocation	Allocation	Allocation	Allocation
S1 W.Score	17.86***						
	(2.132)						
StdDev. Participant's Category Scores	-8.517*	-7.326**	-6.760*	-6.366*	-7.461**		-4.853
	(4.336)	(3.705)	(3.655)	(3.647)	(3.632)		(3.925)
Expert	7.675***	6.396***	6.022**	4.682**	-2.081		-1.961
	(2.801)	(2.383)	(2.330)	(2.344)	(4.138)		(4.215)
S1 Implied Allocation		4.486***	3.006***	2.541***	3.009***	3.008***	2.999***
		(0.332)	(0.555)	(0.565)	(0.549)	(0.654)	(0.644)
Within Participant Expertise					16.94**		16.81**
					(7.183)		(7.248)
Transf.						-1.228	-1.021
						(1.465)	(1.483)
Feas.						2.510*	2.040
						(1.492)	(1.527)
Team						-0.286	-0.487
						(1.305)	(1.282)
Observations	242	242	242	242	242	242	241
R-squared	0.310	0.496	0.536	0.581	0.547	0.522	0.549
Crowd Treat Controls	YES	YES	YES	YES	YES	YES	YES
S1 Rank FE			YES	YES	YES	YES	YES
Presentation Order FE				YES			
Project FE				YES			

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Notes: Appendix Table A.4 reports the results of ordinary least squares regressions where the outcome is evaluator-project level dollar allocation in Survey #2. All models control for peer information treatment status, which captures whether or not the individual was randomized to see their peers' average weighted scores for a given project, and whether or not the peer score was above or below that participant's own weighted score from Survey #1. Column 1 reports the coefficients for the evaluator-project's Survey #1 weighted score, attribute imbalance (standard deviation of the category scores), and a dummy variable for whether the evaluator reported themselves as an "expert" (4/5 or 5/5) on the given project. In Columns 2–7, Survey #1 weighted score is replaced with Survey #1 implied (predicted) allocation, and the r-squared value increases by at least 60% vs. Column 1. Columns 3–7 layer in additional control variables, including evaluator-project session rank fixed effects, project presentation order fixed effects, project fixed effects, evaluator expertise relative to the other projects in the session ($\frac{Expertise_{i,j}}{\sum Expertise_i}$), and specific attribute values.

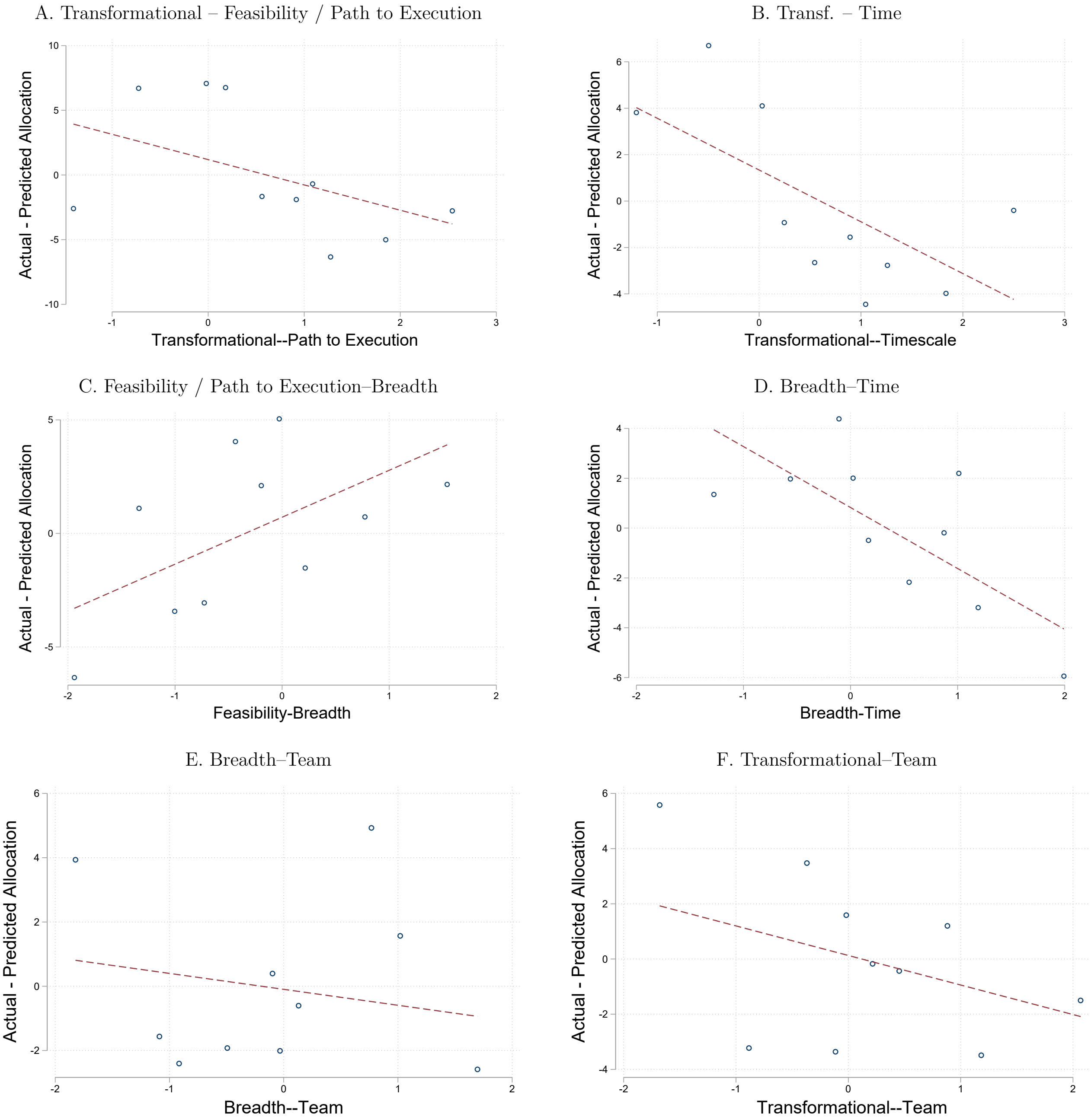
Table A.5: Portfolio Allocation, by Within Evaluation Category Differences

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
VARIABLES	Allocation	Allocation	Allocation	Allocation	Allocation	Allocation	Allocation	Allocation
S1 Implied Allocation	2.770***	2.808***	3.127***	3.167***	2.844***	2.958***	2.977***	3.128***
	(0.553)	(0.554)	(0.558)	(0.552)	(0.559)	(0.562)	(0.558)	(0.554)
Top Attribute (0/1): Transf	-1.407	-1.714						
	(2.727)	(2.742)						
Top Attribute (0/1): Path2Execute	6.015**	4.792*						
	(2.380)	(2.653)						
Top Attribute (0/1): Breadth	1.256	0.668						
	(2.264)	(2.333)						
Top Attribute (0/1): Team	2.944	2.510						
	(2.996)	(3.024)						
Attribute Range		-1.625						
		(1.559)						
Transf.-Path2Exec			-1.866**					
			(0.917)					
Transf.-Time				-2.374**				
				(0.966)				
Path2Exec-Breadth					2.101*			
					(1.066)			
Path2Exec-Team						1.698*		
						(0.998)		
Time-Team							1.801*	
							(1.035)	
Impact-PoS								-1.729***
								(0.630)
Observations	244	244	242	243	243	243	244	242
R-squared	0.522	0.524	0.521	0.527	0.508	0.506	0.509	0.528
Crowd Treat Controls	YES	YES	YES	YES	YES	YES	YES	YES
S1 Rank FE	YES	YES	YES	YES	YES	YES	YES	YES

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Notes: Appendix Table A.5 shows the results of ordinary least squares regressions of participant's Survey #2 project dollar allocations on the implied allocation (based on Survey #1 weighted average scores) and various measures of within-evaluation inconsistency. In Column 1, those measures include four dummy variables for which category score was greatest within a participant-project evaluation. Timescale as the maximum scoring category is the omitted variable. Column 2 additionally controls for the magnitude of the largest attribute pairwise difference. In Columns 3–7 the independent variables are the difference in score for the five least correlated category pairs, Transformational-Feasibility, Transformational-Timescale, Feasibility-Breadth, Feasibility-Team, and Timescale-Team.

Figure A.12: Actual vs. Predicted Allocations, by Attribute Pair Differences



A. Transformational – Feasibility / Path to Execution

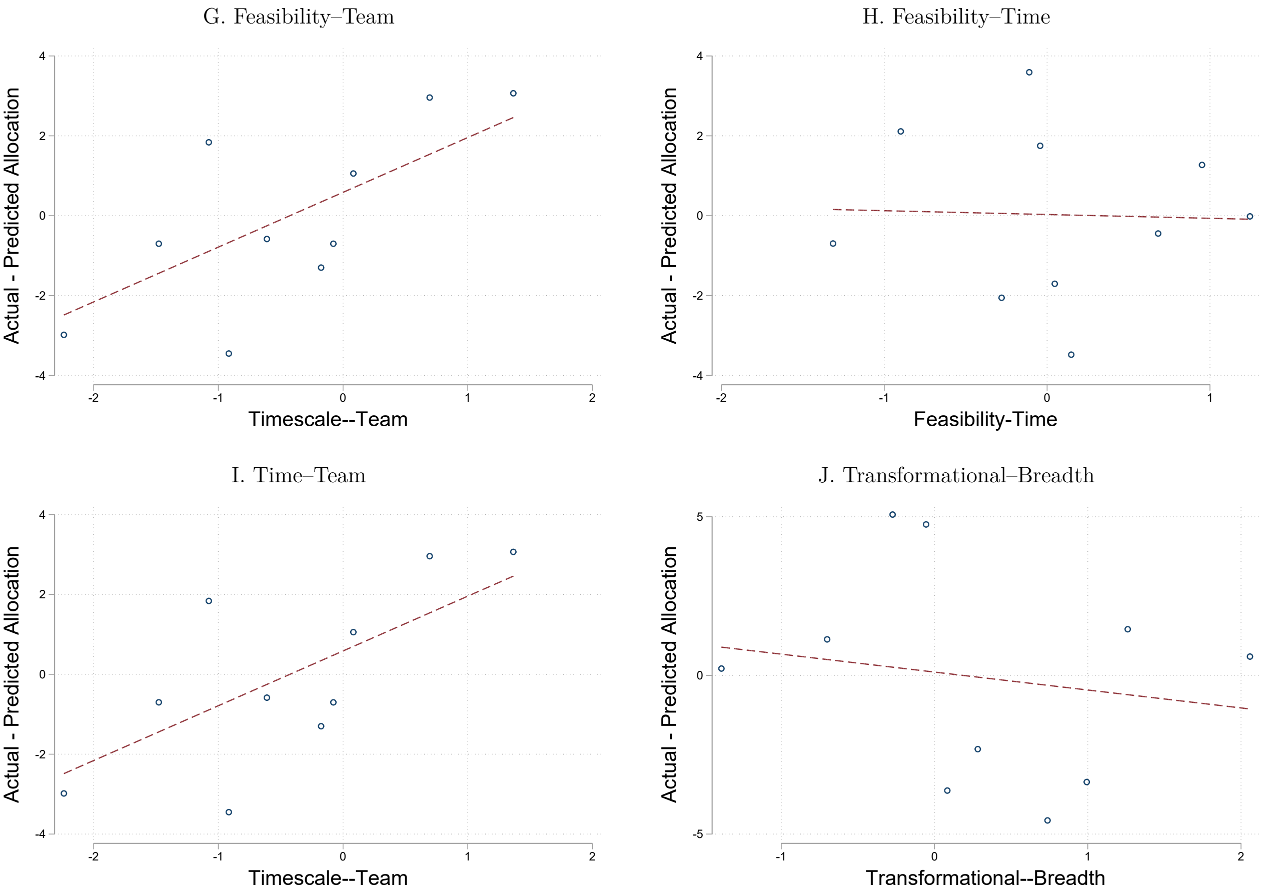
B. Transf. – Time

C. Feasibility / Path to Execution–Breadth

D. Breadth–Time

E. Breadth–Team

F. Transformational–Team



G. Feasibility–Team

H. Feasibility–Time

I. Time–Team

J. Transformational–Breadth

Notes: Figure A.12 shows binscatter plots by 10 quantiles of evaluator-project differences between pairs of category scores. The Y-axis is the difference between participant's actual allocations (Survey #2) and the predicted allocation. All plots adjust for participant-project characteristics (expertise and "home team" bias) and peer score treatment status. The first four plots (A, B, C and D) display the least correlated attribute pairs (i.e., most commonly divergent pairs).