Are Transformational Ideas Harder to Fund? Resource Allocation to R&D Projects at a Global Pharmaceutical Firm[*]

Joshua Krieger
Harvard Business School

Ramana Nanda
Imperial College London

December 28, 2021

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 allocated hypothetical dollars to the same ideas in a head-to-head comparison. We find that in head-to-head comparisons, evaluators systematically penalized projects that they scored high on some attributes but low in others, compared to ones that they scored more equally on all dimensions. A consequence of this pentaly was that ideas perceived as being high risk and high reward tended to lose out to projects considered relatively less transformational, but safer bets. Our results shed light on why novel ideas are handicapped in R&D funding: they may be systematically losing out in head-to-head resource allocation to ideas that are more balanced in their attributes even when the latter have less transformational potential.

JEL Classifications: O31, O32, Q55, L65

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

[*] We are grateful to John Galvin for research assistance with the Pitch Day live surveys. We also thank participants at the 2021 Harvard/Duke/UCSD Workshop on Field Experiments in Strategy, Innovation, and Entrepreneurship, as well as the HBS Entrepreneurship Unit brownbag lunch for helpful comments.

1 Introduction

R&D productivity is on the decline. Many explanations for this decline focus on institutional and external factors that hamper willingness to explore, while steering innovation towards less novel and lower value directions.[1] However, despite resource allocation being such an important part of what all organizations do, scholars have placed less emphasis on *how* organizations solicit, select and fund early stage ideas.

A key challenge in studying resource allocation to early stage ideas is the data: it is hard to access to a risk-set of projects being considered, to understand how they are evaluated along different dimensions and how this maps onto what does and does not get funded. We overcome this challenge by using unique data of actual projects in a real world setting through a collaboration with an internal startup program at the one of the largest pharmaceutical companies in the world. The program, which was set up in order to deliver "transformative, breakthrough innovation" to the parent company screened projects in four stages. After soliciting applications from teams across the organization, a peer-review screening process over three rounds pared down 165 applications to a shortlist of twelve, whose leaders then pitched to a committee of scientific experts responsible for selecting four to fund. This final stage is the context of our study. We partnered with the organizers – who had invited participants from across the organization – to have participants live score projects along each of five dimensions used by the actual evaluation committee, then later allocate hypothetical dollars across projects in each session in a head-to-head comparison.

[1] The decreasing productivity in R&D has been documented across a number of scientific fields and industries (Jones 2009; Bloom et al. 2020; Clancy 2021), and has been an object of particular scrutiny in the pharmaceutical industry (Pammolli et al. 2011; Scannell et al. 2012; Cockburn 2006). Critics claim that pharmaceutical companies shy away from novel high-value R&D in favor of exploiting existing technologies and "me-too" drugs. These critics attribute this lack of innovation to a range of explanations including short-termism, financial constraints, prices not tied to quality, or corruption (Angell 2005; Fojo et al. 2014; Dolgin 2018; Prasad et al. 2018; Coy 2021). The less controversial empirical claim is that novel drug development comes with both higher risk and higher reward (Krieger et al. 2021), and that "fast followers"—whether copy cats or parallel R&D efforts—have moved faster and cut into pioneer profits (DiMasi and Faden 2011; Schulze and Ringel 2013).

Our goal in this paper is to understand resource allocation choices and the degree to which projects with certain combinations of attributes are systematically more or less likely to receive funding. In particular, we try and distinguish between the perceived quality of projects scored independently on several dimensions (a so called “unconstrained” or “independent” scoring regime) and the types of characteristics evaluators choose to fund in a more “constrained” resource allocation environment where there is an ex-ante limit on the number of projects that can be funded, thereby requiring head-to-head comparisons. Our unique setting and two step evaluation process provides a unique window into the values and preferences of the R&D experts who propose and select innovative projects.

Three features of the setting are particularly useful for our analysis. First, R&D personnel from across the organization were invited to view the pitches and given the option to participate in a live survey we conducted of attendees soliciting their evaluation of projects. Although these personnel were not directly responsible for selecting ventures, they had strong and relevant technical backgrounds for understanding and evaluating these ideas (e.g., more than two-thirds are MDs and/or PhDs) as well as institutional knowledge of the organization (average tenure at the firm was 11 years). Altogether, we collected data from 141 participants, resulting in 503 participant-project evaluation pairs across the 12 projects.[2] This 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. This made the information-environment very comparable for the ‘crowd’ evaluators we study relative to the actual evaluation committee. Moreover, the real-time evaluations picked up each individual’s beliefs without any confounding factors

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

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, then later allocate hypothetical dollars across the four projects in each of session. This design allows us to analyze how an individual's relative dollar allocations relate to their own independent project-specific evaluations. This two step measurement enables us to measure the project attributes that the evaluators give greater weight to, independent of what the organization's objectives might be.

Our analysis yields three sets of key findings. First, we find that on average, our set of crowd evaluators score projects in a way that is extremely congruent with other actual evaluators in this program. The pre-pitch scores received by the projects in the final round of peer-evaluation (and unobserved to our evaluators) were strongly correlated with the average scores of the crowd. Moreover, three of the top four projects when ranked by average score were selected for funding by the actual evaluation committee. Given the expertise embedded in the participants who were invited to attend (e.g. over a quarter had served as reviewers for projects like this in the past), this result is not surprising. However we do view it as a "necessary result" as it helps us generalize our subsequent findings to a broader set of evaluators in similar resource allocation settings within corporate R&D.

Second, we document that although the average scores received by projects from the unconstrained scoring regime were closely clustered—the very top and bottom ranked projects only differed by about 20% (4.0/5 vs. 3.3/5)—the *within evaluator attribute imbalance*, or variation in scores across the different attributes received by the same evaluator, was substantial. 50% of project evaluations had a maximum range across categories of 2 or more (on a 5-point Likert scale). Within evaluator attribute imbalance was more prevalent for some projects than others, and was uncorrelated with the project's average score. In other words, some projects were associated with bigger strengths but also more salient weaknesses, while others received a similar average score, but were perceived to be more average across all their attributes. Most notably, projects with higher levels of attribute imbalance were those that tended to consistently score higher than average on impact attributes such "transformational potential" and "breadth," while they were simultaneously scored lower than average on execution attributes like "feasibility" and "timescale to prototype."

Third, we find that despite clear guidelines on how the organizers wanted evaluators to "weigh" different attributes, individuals appeared to systematically emphasize certain attributes more heavily when making head-to-head allocations. Specifically, evaluators appeared more willing to fund a project that was average on all dimensions relative to one with the same average score that scored very high on some measures, but lower in other measures. This inconsistency penalty was driven by projects that scored well on impact measures (transformational potential and breadth), but relatively poorly on execution dimensions (feasibility and timescale), which as noted above are the most common type of attribute imbalance.

A key limitation of our setting is that we do not have data on outcomes, either among those funded or those not funded. Nevertheless, given the generalizability of the context of our study, both in terms of the actual projects being evaluated as well as the scientific qualification and backgrounds of the evaluators we surveyed, our results provide a potential mechanism for the perceived lack of funding for the most transformational ideas when seeking R&D funding: 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 transformational potential. 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. We note that a heuristic that tends to favor consistent within-project qualities and penalize projects with both bigger strengths and more salient weaknesses may be rational and even advisable in certain contexts. However, recent evidence suggests it could be sub-optimal in contexts such as early stage innovation (Malenko et al. 2021). In a related vein, famed venture capital investor, Marc Andresseen has noted that how even the most successful VC investments inevitably have flaws at early stages; focusing on these can lead resource allocators to pick the "mush in the middle" that has no big weaknesses but also no overpowering strengths.

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 run invest in drug candidates targeting diseases that have shorter clinical trials due to surrogate endpoints (Budish et al. 2015). Firms may also choose to limit their R&D differentiation in order to take advantage of the information spillovers of working among competitors with similar technologies (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.

Our work also speaks to the nascent but growing literature on project selection process in science funding, which has largely focused on research grant processes (often at the National Insitutes of Health) and on the composition of the selection committees and applicants (Boudreau et al. 2016; Li 2017; Myers 2020; Azoulay and Li 2020; Lane et al. 2021). Further downsteam, the selection of startups by venture capitalists has also been an area of interest (Kerr et al. 2014; Lerner and Nanda 2020; Ewens et al. 2018), though this research has mainly focused on the consequences of recieiving private capital and 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. Other studies have studied pharmaceutical firm's investment allocation using using retrospective analysis of funded projects and their outcomes (Cook et al. 2014; Morgan et al. 2018; Shih et al. 2018).

2 Research Setting

Our research setting is an internal project funding process at a global pharmaceutical company. Our data comes from the third edition of this program, which the company ran annually starting in 2017. The company initiated this program to surface and fund "transformative breakthrough innovation." The program leadership told us that the science-driven R&D organization was often quick to shoot down high-risk, high-reward project ideas. Most of the firm's scientists came from academic labs, where they were well-trained in "organized skepticism" and valued the skill of poking holes in project ideas. By funding radical projects and pulling their teams out of their usual R&D roles and into an "internal startup" environment, the leadership hoped to discover and commercialize novel ideas, as well as "foster a culture of entrepreneurial thinking to drive cross-disciplinary collaboration."

The winning teams would be given an average of roughly $2 million in special funding and 18 months to work on their project idea. In addition to moving out of their usual research teams into this "startup" environment, the winning teams would be assigned expert project mentors and have access to technologies across the company in order to test their ideas, as well as business support services from legal, information technology, operations, etc.. The focus of the program is on facilitating rapid prototyping for "therapeutics enabling technologies."

The program solicited ideas from R&D teams across more than 20 global research offices and four different research divisions. In particular, the solicitation encouraged proposals that would not fall into the normal project funding pathways in the following areas:

- *New Biology*: novel drug targets and assays
- *New Technologies*: automation, instrumentation, bioinformatics, machine learning, drug creation/delivery
- *Clinical and Translational Science*: patient monitoring, digital disease management tools

2.1 Peer Review Phase

Teams of scientists within the company could submit a short description of their ideas. Teams generally consisted of 3-5 scientists, and often had cross-functional backgrounds and came from multiple offices. After an initial triage process, in which the core program team removed 105 non-viable applications that did not fit in the program, the remaining 55 teams were invited to further develop their project proposals and submit them to a peer review process. 146 company scientists participated in the peer review process of 51 final proposals.[3] 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:[4]

- *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

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

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

- *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.[5] This competitive process ensured that the final Pitch Day projects were already at a high quality baseline, such that a set of experienced company 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.

2.2 Pitch Day

The core team selected 12 proposals to pitch their ideas to a panel of eight company leaders and one external academic expert. 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. Pitchday 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 videoconferencing software. The pitches were organized into three sessions of four projects each (grouped by broad topics). Each team had 10 minutes to present,

[5] To inform our research, we sat in on these selection meetings but did not participate. 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.

followed by up to 10 minutes of questions and answers with the panelists. In between sessions, there was a 15 minute break.

2.3 Data Gathering

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.[6] Ultimately, we had 141 session participants.[7]

Table 1, Panel A reports the composition of the participants. In response to the question "In which field did you earn your highest degree?", 35% responded biology or biochemistry, 32% chemists, 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.

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.2, 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.

[6] Organizers did mention that the evaluation surveys were part of a collaboration with Harvard researchers to study project evaluation.

[7] 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.

2.4 Research Design

The project evaluations were collected in a series of “live” digital surveys that participants filled out throughout the pitch sessions at specified intervals. 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 sequence of session events and examples of the survey interface are depicted in Appendix Figure A.1. Each of four project teams (per session) presented their proposal for 10 minutes, followed by 10 minutes of questions and answer with the selection panel. At the end of the presentation (before the Q&A), the audience evaluators were instructed to fill out their independent evaluations.[8] 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, feasibility, 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). Some participants dropped off at the end of the sesssions and did not complete this second step of the surveys. Therefore, our participant sample size is smaller in the Survey 2 analyses (61 participants, 244 participant-projects) . 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.”

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

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.[9] 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.

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.

3 "Unconstrained" Evaluation of Project Attributes

3.1 Clustered, but Consistent Project Rankings

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, Panel B shows the Survey 1 weighted score results, by project. Appendix Figure A.2 further shows the category-specific scoring distributions. Three basic results about aggregate project scores immediately jump out. First, the average scores align well with

[9] 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.

both the selection panel's chosen winners and the peer review scores. Panel A of Figure 1 shows the 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. Similarly, Panel B 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, crowd 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. Third, we see a fair amount of heterogeneity for any given project's scores, such that the 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.

One concern in any setting with Likert-scoring is that evaluators apply different standards, resulting in a potentially unbalanced mix of "generous" and "harsh" evaluator scores. Panels A of Appendix Figure A.3 shows our estimate of individual participant generosity based on evaluator fixed effects (controlling for project and attribute). Though we see that a small group of evaluators are indeed significantly 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 (see Panel B of Figure A.3).

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 quite normally distributed (see Appendix Figure A.2, 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.2 and Table A.1). 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 R&D 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.1). Overall, evaluator characteristics appear to have a significant but small (in magnitude) impact on project aggregate scores.

3.2 Attribute Imbalance

In our second set of results, we unpack the heterogeneity in evaluator characteristics. We find that certain attributes and certain pairs of attributes are more likely to be associated with overall within-evaluation imbalance.

Along with the variation at the level of project scores, we find substantial across attribute variation within project evaluations. Furthermore, some projects are especially prone to such incongruous attribute evaluations. Figure 2 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 that inconsistency is associated with average project scores. Panel B shows similar variation in average attribute range for projects ranked high and low based on average weighted score. In other words, some projects were consistently inconsistent across attributes, and those might end up anywhere in the average weighted score distribution.

Figure 3 breaks down the within-evaluation conflict by project attribute. Graphing the standard deviation of attribute scores within project-participant evaluation (after controlling for project fixed effects), we see that attribute inconsistency has a strong positive correlation with transformational scores and the opposite relationship with feasibility and timescale. Project breadth and team scores have only weak correlations with overall attribute inconsistency.

In Figure A.4, 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.[10] 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-feasibility) than it is for the group with the lowest (feasibility-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 timing score, 31% with feasibilty, 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 feasibility (41%), followed by time (39%), team (38%), and breadth (37%). Figure A.4, Panel B shows the pairwise difference distributions

[10] We can only speculate about the reasons for different distributions in raw attribute scores (see Panel A of Figure A.2. 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 feasibility and timescale than to team quality.

for the two pairs with the highest and lowest average inconsistency. Unsurprisingly, the attributes pairs that evaluate probability of success (feasibility and timescale) and those that track potential impact (transformational potential and breadth) show the most positive correlations, while the cross between those two groups shows the greatest pairwise dispersion.

4 Head-to-Head 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 Constrained vs. Unconstrained Scoring Preferences. 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 relative project preference differences in the two regimes, 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 evalutor 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 2), and our findings hold.

The second set of deviations relate 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.[11]

Finally, deviations might 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.

4.1 Actual vs. Predicted Allocations

We find evidence of both amplification and reordering of project rank preferences. Panel A of both Figure Figure 4 and Figure A.5 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

[11] 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.

distribution is much wider (see Panel B of Figure A.5), with an interquartile range of \$4.5–\$43.5 (mean=\$25). This "pulling apart" of the predicted allocation is most evident when we graph actual allocations by predicted allocations. We see that the relationship is linear, with a slope greater than 1, and that the effects are symmetric, such that projects below the median see allocations negatively affected by roughly the same proportions as those above the median benefit.[12] Our econometric analysis in Table 2 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).

Figure 4, Panel B shows project preference reordering from Survey 1 to Survey 2. We find that about 45% 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 10% move more than one full rank (e.g., 1st to 3rd).[13] 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 results in different signals about relative project values.

We also find that the expertise effects are further exaggerated in the portfolio allocation choices. As described in Section 3, expert and home-team effects are already reflected in the independent weighted scores, and therefore in the predicted allocations. We find that project expertise wins another \$4.7–\$7.7 dollars in actual allocations, and that "within participant expertise" (participant's project expertise relative to their total expertise on all projects in

[12] Figure A.6, Panel A plots the 20 quantile binscatter relationship between predicted allocation and the difference between actual and predicted allocation, after controlling for a number of evaluation characteristics (evaluator expertise, home-team effect, 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 projets below and above the median in predicted allocations.

[13] Panel B of Figure A.6 shows that that expected (absolute) rank change is decreasing in predicted allocation, after controlling for evaluation characteristics. The relationship is statistically significant at the 10% level in analogous regression results (not displayed). This pattern suggest that the very top scoring projects from Survey 1 are less likely to change rank in the portfolio allocation evaluation regime.

that session) is also significantly associated with greater dollar allocations.

However, we find that within-evaluation attribute inconsistency is associated with an allocation penalty. The binscatter plots in Panels C and D of Figure 4 show that even after controlling for project fixed effects and evaluator-project expertise, attribute inconsistency—measured as the standard deviation in attribute scores within a given participant-project evaluation—is penalized in the Survey 2 allocations. This relationship is true both at the evaluator level (Panel C) and when we aggregate up to the project level and look at the percentage of a project's Survey 1 evaluations with high attribute inconsistency (Panel D). Our regressions (Table 2) confirm that the negative relationship between attribute inconsistency and the allocation penalty is statistically significant, and holds up as we control for predicted allocation, expertise, Survey 1 rank and presentation order, as well as project fixed effects.

4.2 High Risk, High Impact Projects Drive Inconsistency Penalty

Finally, we drill deeper into which types of projects drive the attribute inconsistency penalty. Our data consistently points to projects that have relatively high potential impact, but also higher execution risk as the projects which experience the largest penalties.

As we discussed in Section 3, participants had substantial variation in their weighted project scores as well as in their within-evaluation attribute pairs, even after normalizing by category and evaluator fixed effects. These patterns imply that the attribute inconsistency penalty is driven by projects scoring high on transformational potential, but low on feasibility and timescale. From our analysis of attribute pairs (see Figure A.4), we know that transformational-feasibility and transformational-timescale differences are also the most likely source of within-evaluation attribute divergence. We verify this implication in the regression analyses reported in Table 3. The first two columns show that having transformational potential or breadth of impact score as the evaluation's highest (normalized) attribute score is associated with a

\$5.89–\$9.24 decrease in Survey 2 allocations, after adjusting for predicted allocations.[14]

Last, we look at how specific pairwise attribute differences are penalized or rewarded in the constrained portfolio allocations. Figure 5 shows the relationship two of the most frequently incongruent attribute pairs (transformational–feasibilty, and feasibility–breadth) and the deviation from predicted allocation. In the binscatter plots, which control for Survey 1 evaluator-project rank (within the session) and evaluator expertise, we see a premium put on feasibility and a penalty for projects that are high in transformational potential and breadth, but lower in feasibility. Figure A.7 shows the same plot for the rest of the attribute pairs. The general pattern holds: a penalty for projects with high potential impact and high execution risk, which a premium for safer, but lower potential impact projects.

Columns 3–7 of Table 3 show these same relationships in regression form for the top five least correlated attribute pairs affect actual allocations. After controlling for overall and relative quality of the project (predicted allocation and Survey 1 rank order), projects which score poorly on feasibility and/or timescale, but relatively well on other dimensions suffer allocation penalties. The strongest penalty appears to be the gap between (normalized) transformational potential and timescale. The results suggest that an evaluator who gave a project a 5 on transformational potential and a 1 on timescale, would—above and beyond their predicted allocation based on weighted Survey 1 scores—shave off an additional \$9.50 off that participant's allocation for that project. That \$9.50 is equivalent to a 38% drop for an average rated project. Even just a 1 point difference in transformational potential and timescale translates to a 9.5% penalty for the average project. Mechanically, the reverse relationship is also true. Projects that excel on the feasibility and timescale dimensions, but less so on other attributes, get a boost in dollar allocation decisions. However, we know from Figure 3 that high inconsistency evaluations are most likely have high transformational,

[14] Transformational potential and breadth of impact are the second most correlated attribute pair, following feasibility-time.

breadth and team scores, and lower feasibility and timescale scores.

In summary, our results show that a specific flavor of divergence drives the inconsistency penalty. Evaluator's portfolio allocations signal a distaste for novelty and potential impact when those characteristics comes at a cost of increased execution risk—and they often do. Since the combination of high transformational potential and low feasibility/timescale are the most common pairs driving attribute inconsistency, that distaste mostly impedes highly transformational projects.

5 Discussion & Conclusion

Our analysis of the two Pitch Day scoring surveys provides a window into the distribution of project qualities as well as evaluators' revealed preferences. We find that R&D professionals at the company provided average project scores that were highly consistent with expert peer reviewers and the selection panel. However despite this consensus (on average), we find substantial heterogeneity in the level and combination of project scores, both across evaluators and within evaluations. The within evaluation attribute imbalance is largely driven by evaluations that see great impact potential (transformational and breadth), but lower likelihood of success (feasibility and timescale). 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 projects with that attribute imbalance.

Our results show how portfolio allocation processes is a chance for evaluators to impose their own values/preferences over the preferred bundles of attributes. And even in the presence of strong guidelines on what types of projects the firm desires (e.g., the instructions signaling 3x weight for transformational potential) scientists may revert to "organized skepticism" and have a hard time seeing past their concerns about a project's hypotheses and execution. If this pattern is present in our setting then we might expect it to be even more prevalent in more traditional project funding committees.

It is notable that the selection panel appeared less prone to the attribute inconsistency bias, as they ultimately did select one project that was below the mean in overall score, but in the top three for transformative potential. However, even if the leadership committee does better internalize the relative importance of impact-based attributes when investing in innovation (essentially call options with greater variance), the results reveal that the R&D rank and file scientists—who are the same people who have to put forward initial project proposals—down-weight impact over probability of success. That they 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 will inhibit high-variance exploration, even when the company signals an appetite to pursue such projects.

References

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

Angell, M. (2005). *The truth about the drug companies: How they deceive us and what to do about it.* Random House Incorporated.

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

Bloom, N., C. I. Jones, J. Van Reenen, and M. Webb (2020). Are ideas getting harder to find? *American Economic Review 110*(4), 1104–44.

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.

Clancy, M. (2021). Innovation gets (mostly) harder: Micro and macro evidence on the productivity of r&d over time.

Cockburn, I. M. (2006). Is the pharmaceutical industry in a productivity crisis? *Innovation policy and the economy 7*, 1–32.

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.

Coy, P. (2021). Out of grief, mit's andrew lo invented a better way to finance biomedical innovation.

DiMasi, J. A. and L. B. Faden (2011). Competitiveness in follow-on drug r&d: a race or imitation? *Nature Reviews Drug Discovery 10*(1), 23–27.

Dolgin, E. (2018). Bringing down the cost of cancer treatment. *Nature 555*(7695).

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.

Fojo, T., S. Mailankody, and A. Lo (2014). Unintended consequences of expensive cancer therapeutics—the pursuit of marginal indications and a me-too mentality that stifles innovation and creativity: the john conley lecture. *JAMA Otolaryngology–Head & Neck Surgery 140*(12), 1225–1236.

Jones, B. F. (2009). The burden of knowledge and the "death of the renaissance man": Is innovation getting harder? *The Review of Economic Studies 76*(1), 283–317.

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*.

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.

Pammolli, F., L. Magazzini, and M. Riccaboni (2011). The productivity crisis in pharmaceutical r&d. *Nature reviews Drug discovery 10*(6), 428–438.

Prasad, V., C. McCabe, and S. Mailankody (2018). Low-value approvals and high prices might incentivize ineffective drug development. *Nature Reviews Clinical Oncology 15*(7), 399–400.

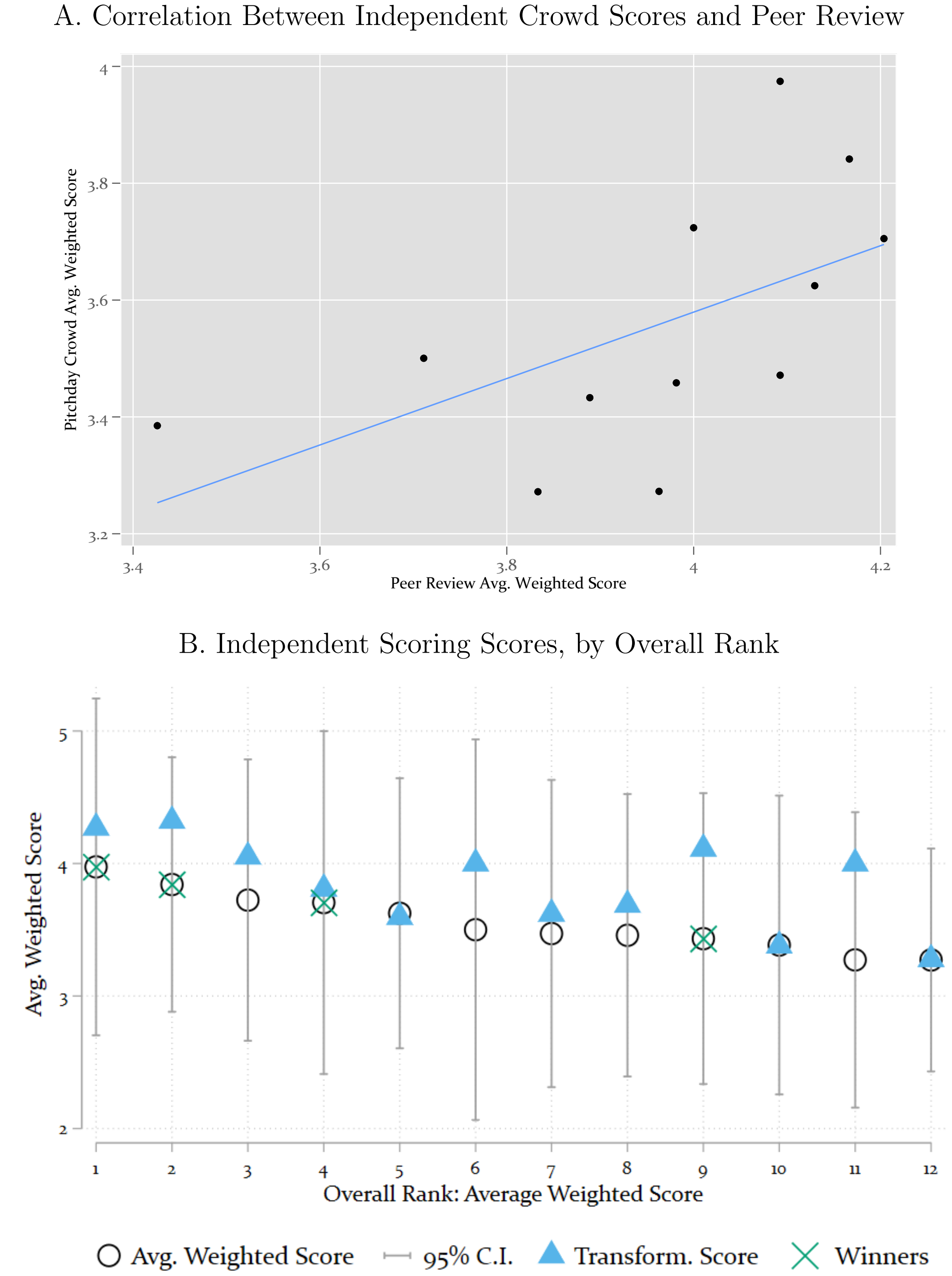
Scannell, J. W., A. Blanckley, H. Boldon, and B. Warrington (2012). Diagnosing the decline in pharmaceutical r&d efficiency. *Nature reviews Drug discovery 11*(3), 191–200.

Schulze, U. and M. Ringel (2013). What matters most in commercial success: first-in-class or best-in-class? *Nature reviews. Drug discovery 12*(6), 419.

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: Independent Scoring: Outcomes and Alignment with Peer Review and Winners

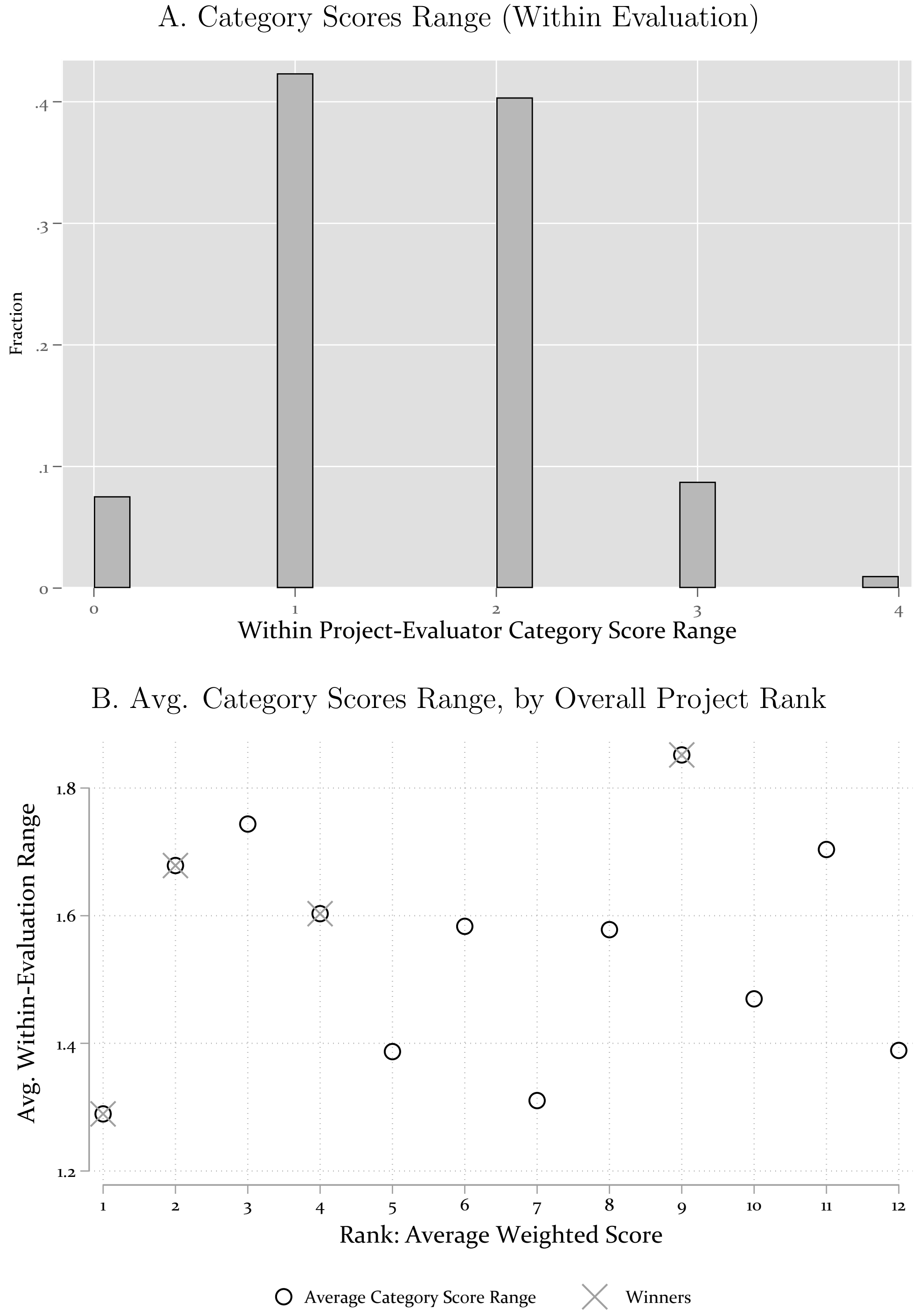


A. Correlation Between Independent Crowd Scores and Peer Review

B. Independent Scoring Scores, by Overall Rank

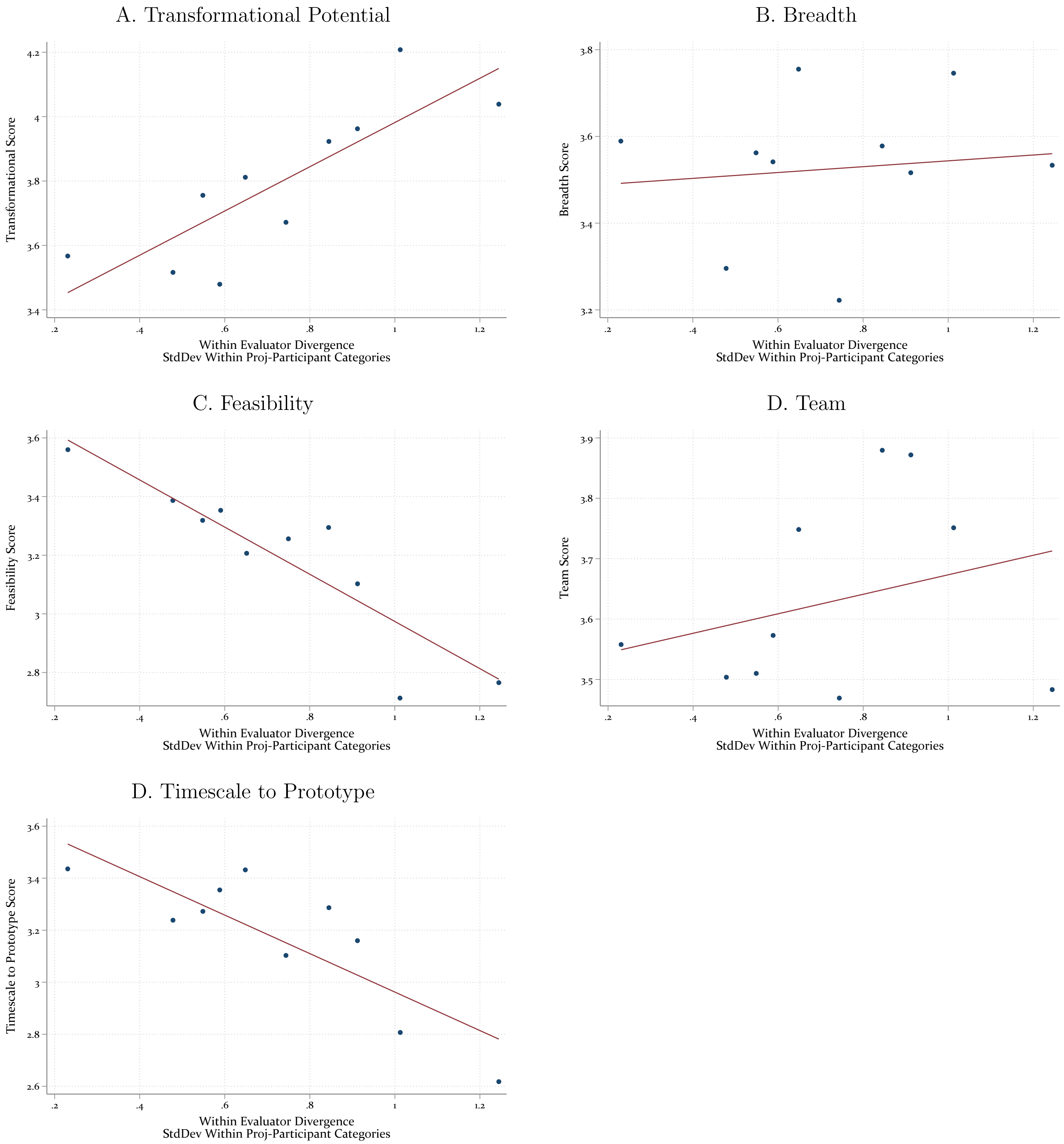
Notes: Figure 1 graphs the Survey 1 (independent category scoring) average weighted scores by projects based on the sample of 141 participants and 503 participant-project evaluations. The weighted scores give extra weight to certain attribute scores (transformational potential: 3x, feasibility: 2x, team: 2x). We order them by rank (best to worst average scores), and display the 90% confidence interval (grey bars). The blue triangles report the average transformational potential score corresponding to each project. The winning projects chosen by the selection panel are marked by Xs.

Figure 2: Range Across Evaluation Categories (Within Evaluation)



Notes: Figure 2 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 3: Figure 3: Conflicted Evaluations and Project Category Scores



A. Transformational Potential

B. Breadth

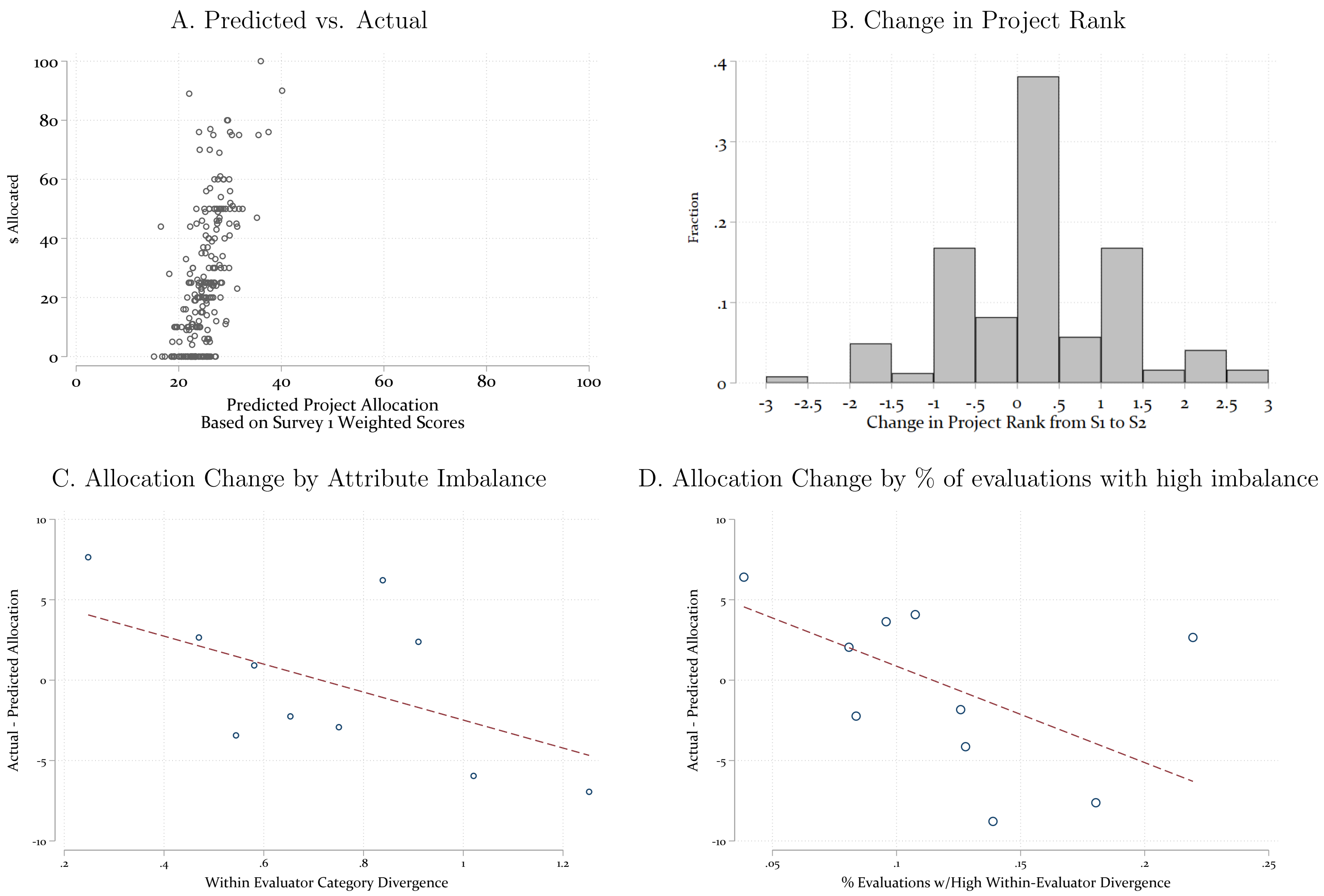
C. Feasibility

D. Team

D. Timescale to Prototype

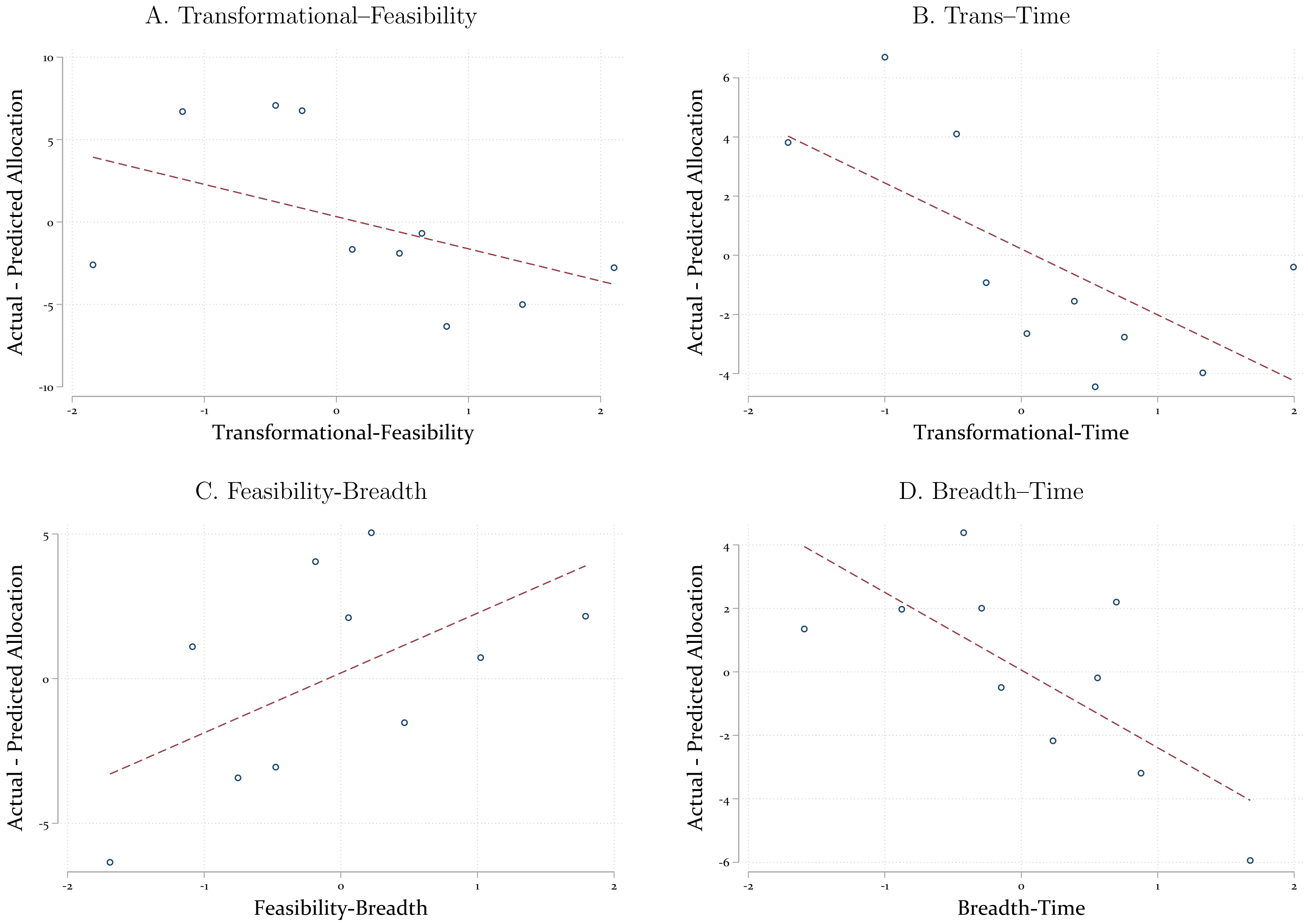
Notes: Figure 3 show the correlation between an evaluation's overall attribute imbalance—the standard deviation across a given participant-project's five category scores—and each one of those scores. Each panel displays the correlations as binscatters by decile of attribute imbalance, after adjusting for project fixed effects and evaluator-project characteristics (expertise and "home team" bias).

Figure 4: Portfolio Evaluation (Survey 2) Predicted vs. Actual Scores



Notes: Figure 4 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 forthe 61 participants (244 participant-projects) completed in Survey 2. Panels C and D show the binscatter plot of the change from predicted allocation (based on Survey 1 scores) to actual project allocation in Survey 2 at the participant project-level. Panel C displays results by quantile of participant-project attribute inconsistency (standard deviation of attribute scores for a given project evaluation). Panel D calculates quantiles based on the percentage of that project's reviews that had high within-evaluator divergence (a range greater than or equal to two Likert scale points). Both binscatters control for project fixed effects, as well as participant-project expertise and "home team" bias.

Figure 5: ALLOCATION PENALTY BY ATTRIBUTE DIFFERENCES



NOTES: Figure 5 shows binscatter plots of the change from predicted allocation (based on Survey 1 scores) to actual project allocation in Survey 2 at the participant project-level. Each of the four plots adjust for project session rank, evaluator expertise, and controlling for peer score treatment. The binscatter plots with the remaining six attribute pairs are displayed in Appendix Figure A.7.

Table 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 1 presents summary statistics from the sample of 141 participants in Survey 1.

Table 2: Portfolio Allocation, by Predicted Values and Evaluation Characteristics

VARIABLES	(1) Allocation	(2) Allocation	(3) Allocation	(4) Allocation	(5) Allocation	(6) Allocation	(7) 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 2 reports the results of ordinary least squares regressions where the outcome is evaluator-project level dollar allocation in Survey 2. All models control for crowd 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 3: Portfolio Allocation, by Within Evaluation Category Differences

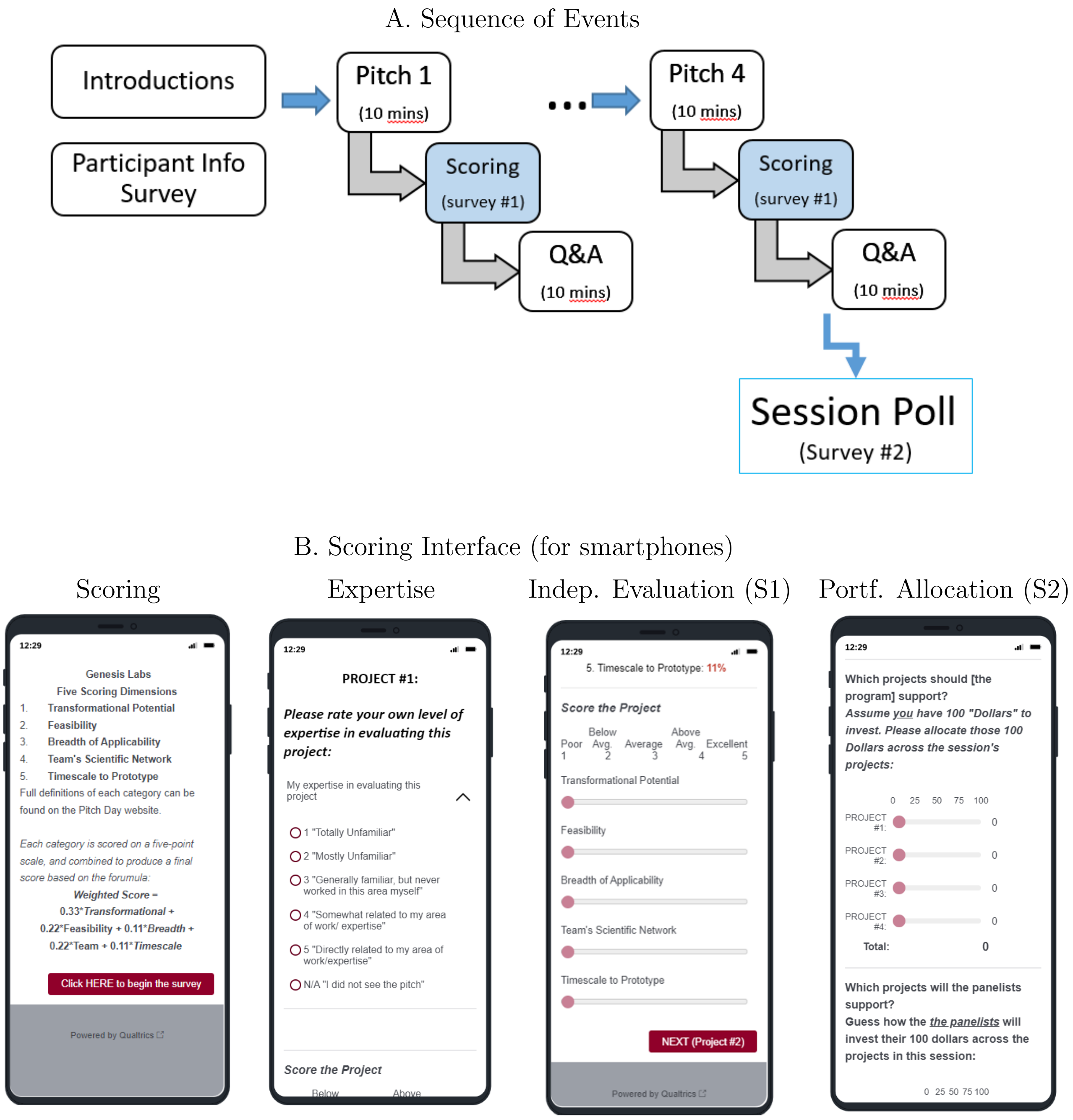
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
VARIABLES	Allocation	Allocation	Allocation	Allocation	Allocation	Allocation	Allocation
S1 Implied Allocation	2.834***	2.845***	3.127***	3.167***	2.844***	2.958***	2.977***
	(0.548)	(0.550)	(0.558)	(0.552)	(0.559)	(0.562)	(0.558)
Top Attribute (0/1): Transf	-6.206**	-5.891**					
	(2.709)	(2.790)					
Top Attribute (0/1): Feas	-2.467	-2.507					
	(3.539)	(3.546)					
Top Attribute (0/1): Breadth	-9.010**	-9.242**					
	(4.159)	(4.194)					
Top Attribute (0/1): Team	-2.067	-2.567					
	(3.455)	(3.611)					
Attribute Range		-0.801					
		(1.652)					
trans-feas			-1.866**				
			(0.917)				
trans-time				-2.374**			
				(0.966)			
feas-breadth					2.101*		
					(1.066)		
feas-team						1.698*	
						(0.998)	
time-team							1.801*
							(1.035)
Observations	244	244	242	243	243	243	244
R-squared	0.526	0.526	0.521	0.527	0.508	0.506	0.509
Crowd Treat Controls	YES	YES	YES	YES	YES	YES	YES
S1 Rank FE	YES	YES	YES	YES	YES	YES	YES

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

Notes: Appendix Table 3 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 and evaluator normalized) score was greatest within a participant-project evaluation. Timescale is the omitted variable, and we include the maximum pairwise gap (between normalized category scores) as an additional control 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 (normalized) score for the five least correlated category pairs, transformational-feasibility, transformational-timescale, feasibility-breadth, feasibility-team, and timescale-team.

APPENDIX

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



A. Sequence of Events

B. Scoring Interface (for smartphones)

Scoring

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

Expertise

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

Indep. Evaluation (S1)

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

Portf. Allocation (S2)

12:29

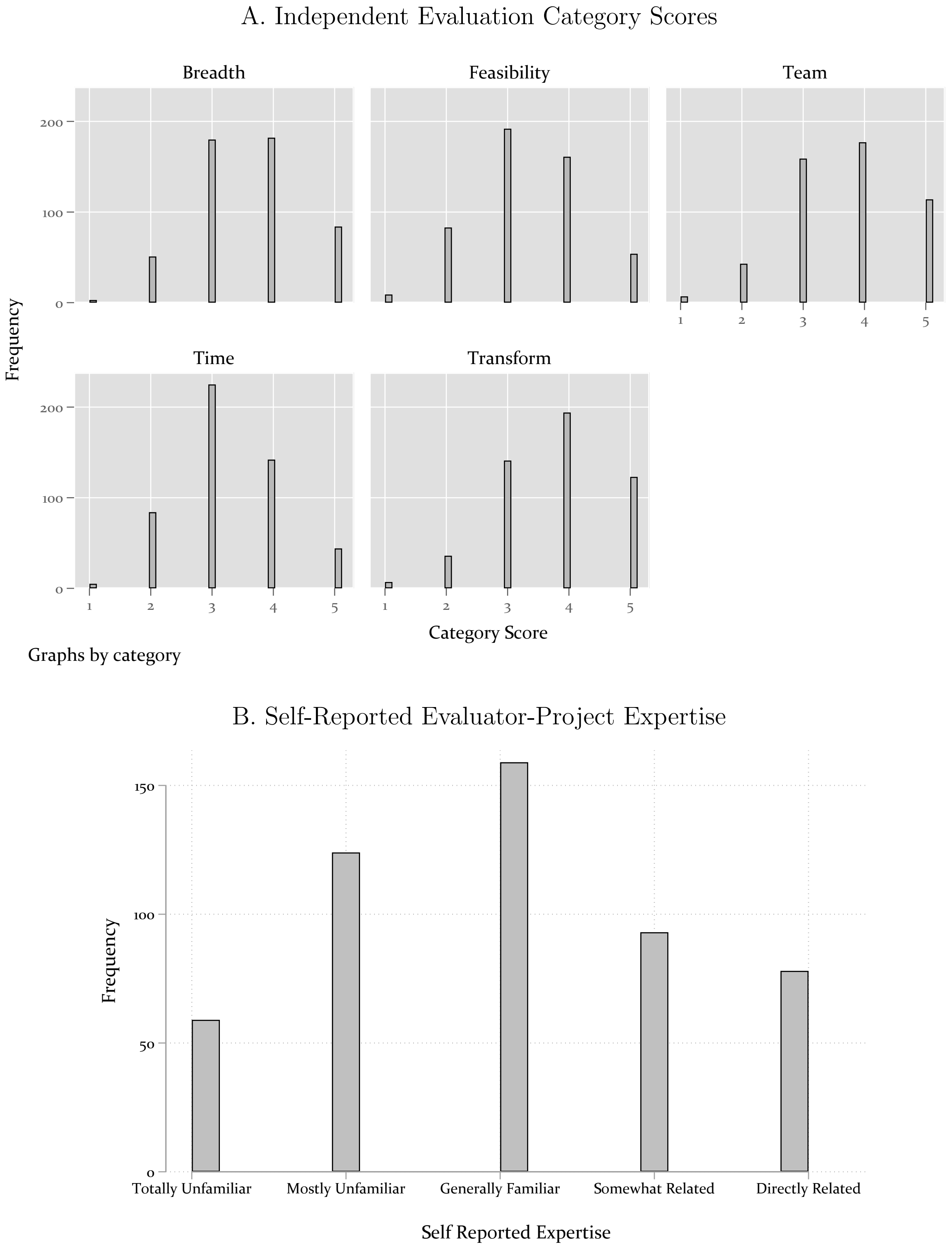
Which projects should [the program] support?
Assume you have 100 "Dollars" to invest. Please allocate those 100 Dollars across the session's projects:

	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: Distribution of Project-Category Scores (Survey 1)

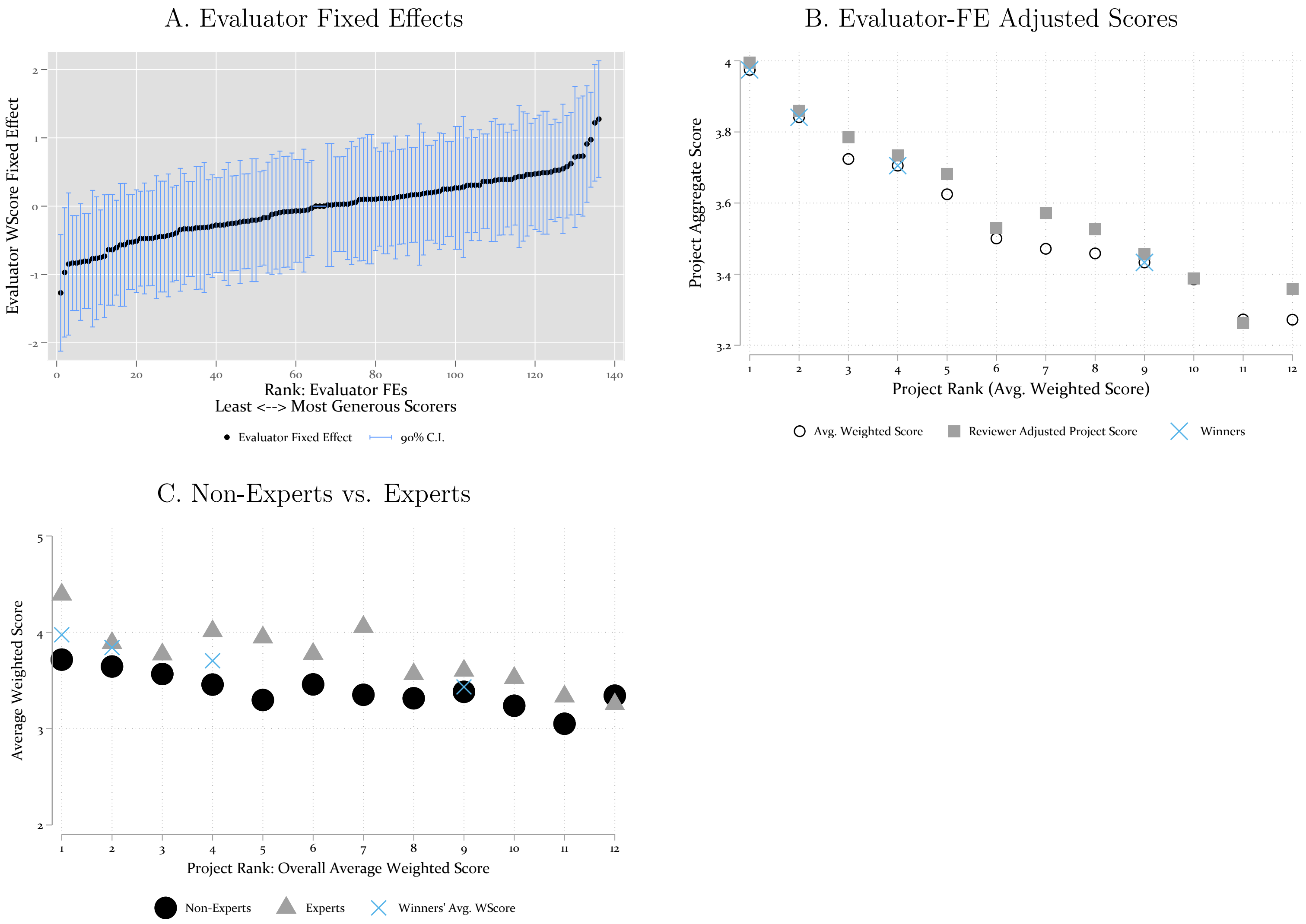


A. Independent Evaluation Category Scores

B. Self-Reported Evaluator-Project Expertise

Notes: Figure A.2 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.3: Independent Scoring (Survey 1) Outcomes: Evaluator and Expertise Effects



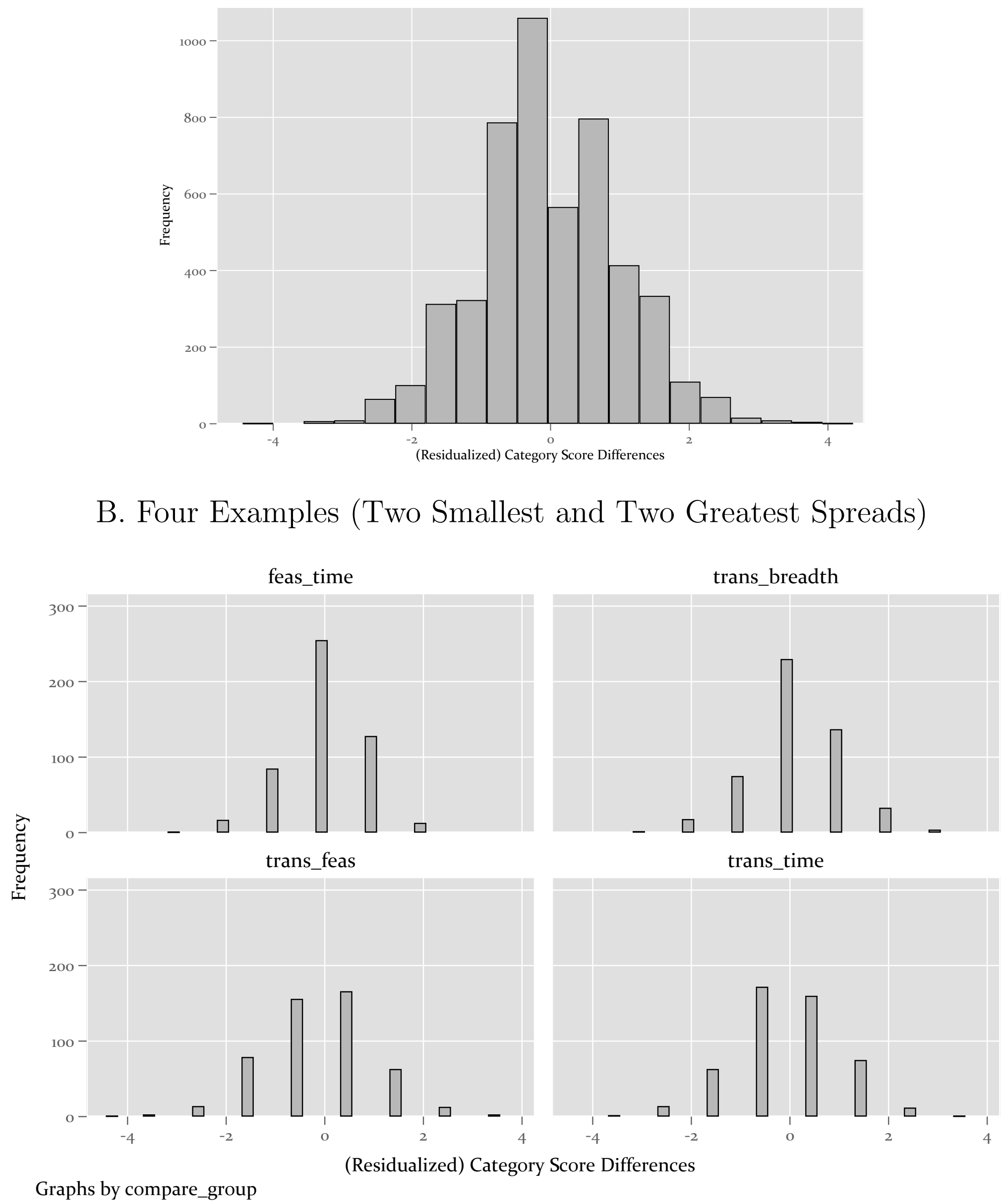
A. Evaluator Fixed Effects

B. Evaluator-FE Adjusted Scores

C. Non-Experts vs. Experts

Notes: Figure A.3, 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.

Figure A.4: Within-Evaluation Pairwise Differences

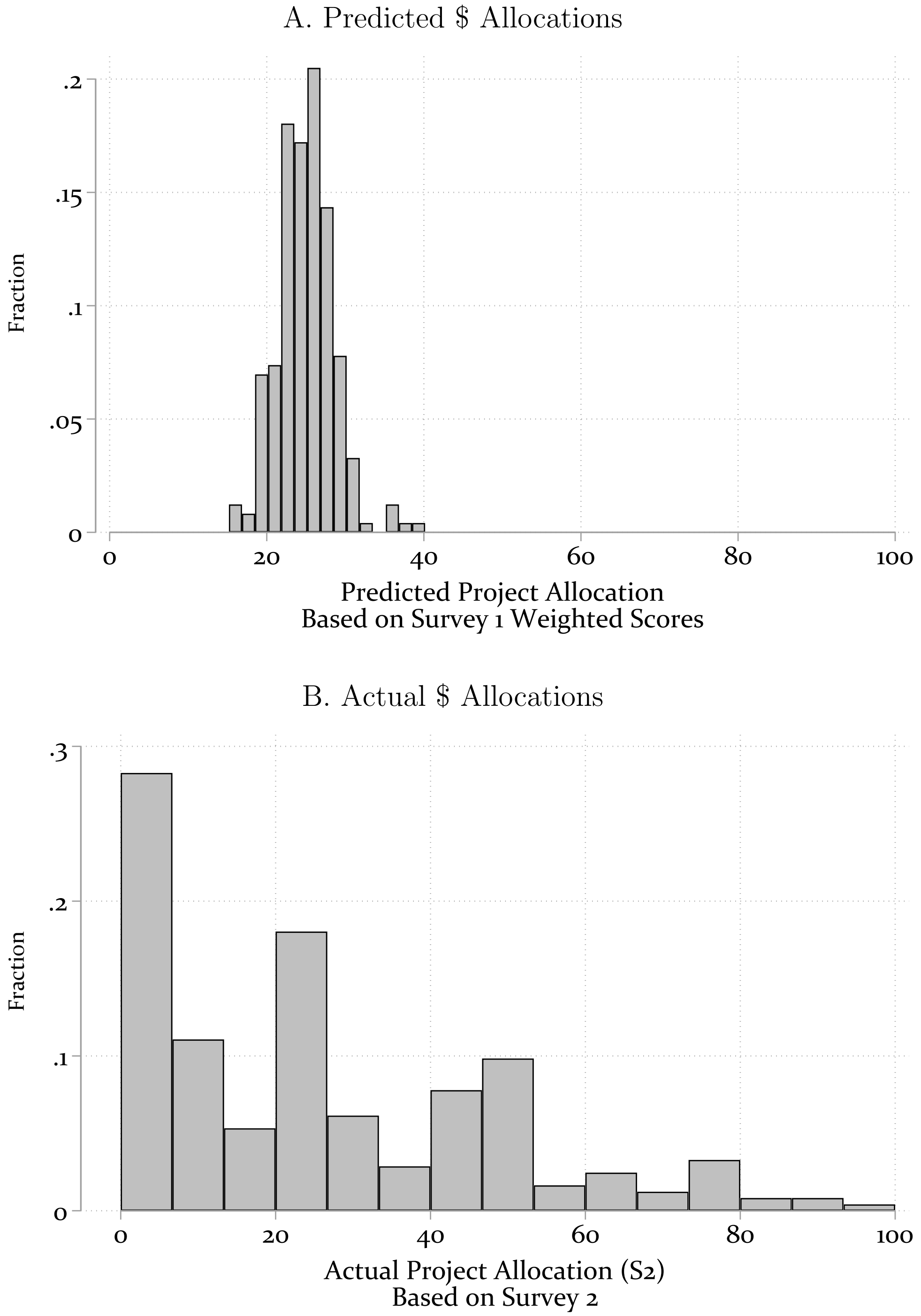


A. Residualized Attribute Differences

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

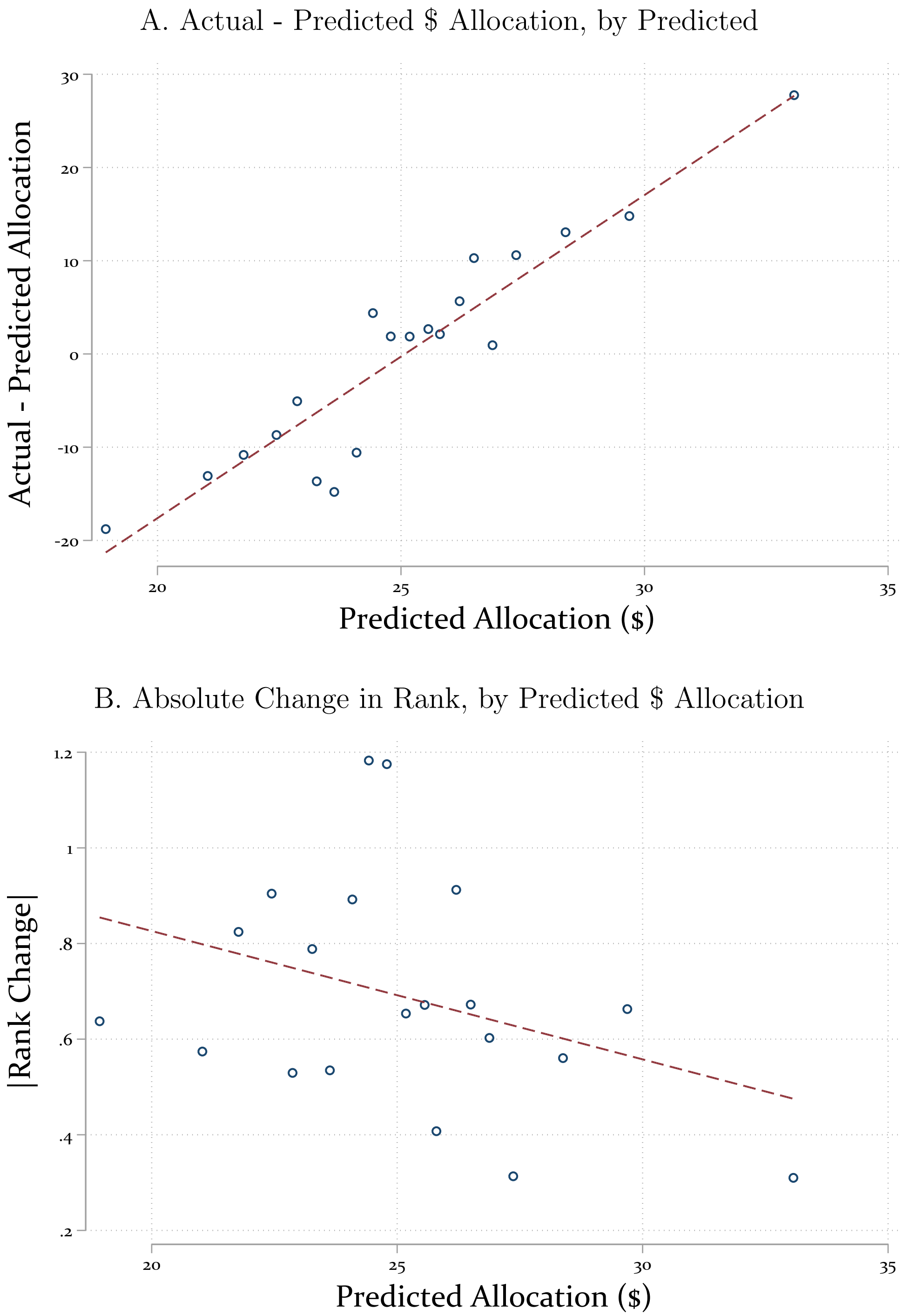
Notes: Figure A.4 shows the histograms category score differences within participant-project-evaluations. The score are first normalized to account for category and evaluator fixed effects. Next, we calculate all pairwise differences between the (residualized) category scores within a participant-evaluation. 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.5: Portfolio Evaluation (Survey 2) Predicted vs. Actual Allocation Distribution



Notes: Figure A.5 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.6: Actual vs. Predicted Allocations and Ranks (S1 vs. S2), Binscatters

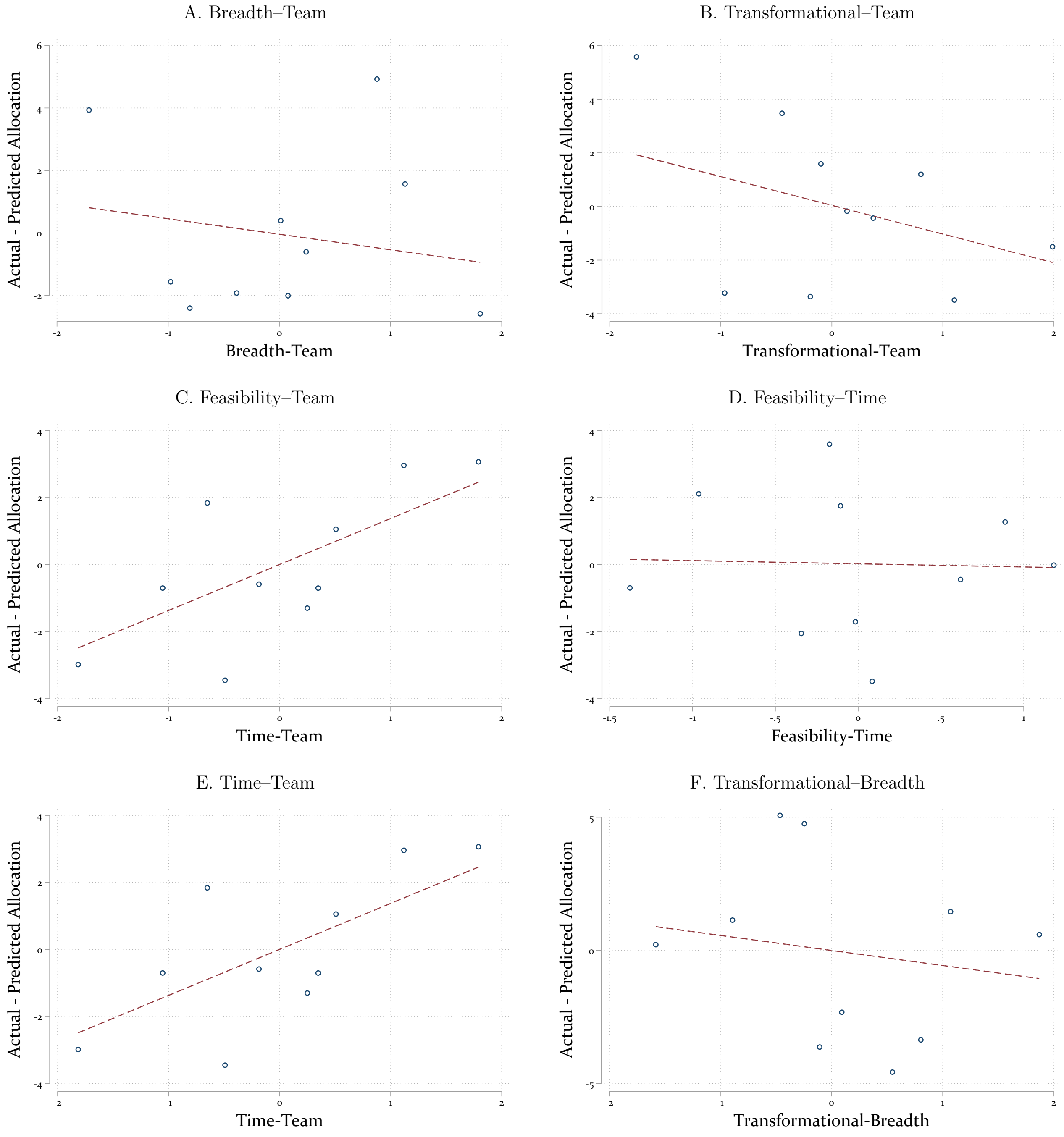


A. Actual - Predicted $ Allocation, by Predicted

B. Absolute Change in Rank, by Predicted $ Allocation

Notes: Figure A.6 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.

Figure A.7: Actual vs. Predicted Allocations, by Attribute Pair Differences


A. Breadth–Team

B. Transformational–Team

C. Feasibility–Team

D. Feasibility–Time

E. Time–Team

F. Transformational–Breadth

Notes: Figure A.7

Table A.1: 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.1 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 variabel 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.