NBER WORKING PAPER SERIES

UNCERTAINTY AND INDIVIDUAL DISCRETION IN ALLOCATING RESEARCH FUNDS

Anna Goldstein
Michael Kearney

Working Paper 32033
http://www.nber.org/papers/w32033

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
January 2024

Our analysis originated as a consulting engagement with the National Academies of Science, Engineering and Medicine for a study on ARPA-E (National Academies, 2017). A. P. G. was supported by a fellowship from the Belfer Center for Science and International Affairs. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors. We are thankful for helpful discussions with Laura Diaz Anadon, Pierre Azoulay, Paul Beaton, Iain Cockburn, Gail Cohen, Jeff Furman, Daniel Kim, Josh Krieger, Gilbert Metcalf, Ramana Nanda, Venky Narayanamurti, Scott Stern, and participants in the NBER Productivity and Innovation seminar. We also thank ARPA-E staff who assisted with data collection, in particular Dave Dixon, Ron Faibish, Andy Kim and Ashley Leasure. All errors or omissions are our own. Michael Kearney is an investor with The Engine Ventures, a firm that supports some companies funded by ARPA-E, the empirical focus of this paper. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Uncertainty and Individual Discretion in Allocating Research Funds
Anna Goldstein and Michael Kearney
NBER Working Paper No. 32033
January 2024
JEL No. O31,O38

ABSTRACT

There is a long-standing tradition in public research funding agencies of distributing funds via peer review, which aggregates evaluations of proposed research ideas from a group of external experts. Despite complaints that this process is biased against novel ideas, there is poor understanding of an alternative system that may overcome this bias: the use of individual discretion. Here, we conduct the first quantitative study of how individual discretion affects a research funding portfolio. Using internal project selection data from the Advanced Research Projects Agency-Energy (ARPA-E), we describe how a portfolio of projects selected by individual discretion differs from a portfolio of projects selected by traditional peer review. We show that ARPA-E program directors tend to fund proposals with greater disagreement among experts, and they also appear to prefer proposals described in reviewer comments as "creative." These choices do not result in a significant tradeoff with short-term project performance, and they enable ARPA-E to fund more uncertain and creative research ideas, which supports the agency's mission of pursuing novel ideas for transformational energy technology.

Anna Goldstein
Prime Coalition
apgoldst@gmail.com

Michael Kearney
Massachusetts Institute of Technology
mkearney@mit.edu

A data appendix is available at http://www.nber.org/data-appendix/w32033

1. Introduction

There exists broad scholarly agreement on the importance of public research funding for promoting innovation and economic growth, due to the underinvestment of the private sector and the cumulative nature of knowledge (Arrow, 1962; Dasgupta and David, 1994). Despite the strong justification for public research funding organizations, there has been insufficient empirical review of how such programs are managed and, in particular, how they decide to allocate their funding across proposed research projects. The key challenge to project selection is dealing with the inherent uncertainty of innovation. There is significant uncertainty both in the research itself and in the "realm of human activity," i.e. markets for technology (Rosenberg, 1996, 1990). Together, these uncertainties make it extremely difficult to forecast the value of any research activity *ex ante*.

In public research funding agencies, the dominant project selection method is peer review, wherein projects are selected by aggregating the opinions of a group of external experts. The drawbacks of peer review have been frequently articulated in the literature. Among these are complaints that reviewers discount novel ideas (Boudreau et al., 2016) and that consensus decisions result in a failure to fund "high-risk/high-return research" (Linton, 2016).

Meanwhile, empirical assessment of alternative designs for research funding allocation has been elusive. In particular, there is a gap in the literature concerning the method of project selection that empowers program staff to use individual discretion. The Defense Advanced Research Projects Agency (DARPA) is the canonical example of an agency that gives program managers freedom to distribute funds without soliciting external peer review. Research on DARPA has focused primarily on active program management, wherein autonomy in project selection is but one element of a larger programmatic strategy (Bonvillian and van Atta, 2011; Fuchs, 2010), and has been mostly limited to qualitative studies.

In this paper, we add to this literature with a quantitative assessment of project selection practices at the Advanced Research Projects Agency-Energy (ARPA-E): an agency that, like DARPA, bestows its program staff with discretion to make funding decisions. ARPA-E provides a unique opportunity for this assessment because, while they give program directors discretion in selecting proposals to fund, they also solicit peer reviews of all submitted proposals. This allows us to compare ARPA-E's research portfolio with counterfactual alternative portfolios and address the following questions: How does a portfolio of projects selected by individual discretion differ from a portfolio of projects selected by traditional peer review? How do these decisions relate to the opinions of external peer reviewers? And what is the impact of this project selection practice on the research outputs of the portfolio?

We answer these questions using data on research proposals submitted to ARPA-E from 2009-2015 and interviews with multiple program directors. Our findings are five-fold: (1) approximately half of ARPA-E projects are "promoted", i.e. selected despite low review scores; (2) proposals are more likely to be selected if reviewers disagree on the quality of the proposal, particularly if the proposal has at least one champion; (3) reviewer comments likely play a role in project selection, as certain words, e.g. *creative*, are correlated with selection; (4) idiosyncratic differences between program directors and the need to construct a diversified research portfolio may also influence project selection; and (5) "promoted" projects perform equally well on average compared to non-"promoted" projects on short-term metrics.

2. Background

2.1. Decision-making for research funding

The conventional approach to allocating public research funds has long been to use a peer review process of some kind to determine which research ideas are worth funding. For nearly as long, there have been debates within the scientific community about the pitfalls of peer review. Many have criticized peer review for its inefficiency and its conservative bias, while others defend peer review for its resistance to corruption and political influence. In this section, we review some of the broad discussion around peer review, in order to compare with the use of individual discretion.

The origin of modern peer review for funding proposals in the US has been traced to the 1940's in the Office of Naval Research, with "an informal 'seeking of a second opinion' by the grants manager, who mailed a copy of a proposal on the periphery of his competence to a colleague and followed up with a phone call" (Roy, 1985). This method of gathering outside opinions as decision-making inputs gained popularity, and a variety of different peer review systems have since proliferated, perhaps due to the appeal of a "system of institutionalized vigilance" that could produce an impartial appraisal of an idea's merit (Merton, 1973). Over time, political pressure on US federal agencies has put more decision-making power in the hands of reviewers rather than program staff (Baldwin, 2017).

The most common implementation of peer review for grant-making, labeled "traditional peer review" by Guthrie and co-authors (2013), is as follows: a set of proposed projects are evaluated by a group of experts, either as an in-person panel or individually in writing. Panelists may be asked to reach consensus on which proposals should be funded, or they may simply be asked to submit their individual opinions after discussion. In either case, proposals are typically ranked in order of funding priority and some portion of proposals is funded, depending on the budget of the program. This generic description applies

to the peer review process at many grant-making organizations—most notably the National Institutes of Health (NIH), which collectively entail the largest public research investment in the US with a budget of $32 billion in 2016 (National Institutes of Health, 2017).

A vast literature outlines the many potential shortcomings of peer review for research proposals. Generally, complaints relate to the equity, efficiency, or effectiveness of the process (Guthrie et al., 2013; Ismail et al., 2009; Wessely, 1998). Regarding equity, Lee et al. (2013) reviews commonly cited sources of bias in peer review including, but not limited to, nationality, language, gender, and prestige. Others have highlighted the perception that insular "old boy" networks contaminate peer review systems (Gillespie et al., 1985). There are also concerns regarding the efficiency of peer review, i.e. whether the process can be administered at a reasonable cost (Gordon and Poulin, 2009). As a smaller proportion of proposals receive funding, the time burden on individual researchers is greater. Researchers must prepare and submit more proposals, while also spending more time reviewing others' proposals (National Institutes of Health, 2008).

As for effectiveness, there is a stream of literature investigating the ability of peer review to identify the best projects *ex ante*.[1] Li and Agha (2015) found value in peer review for nearly 30 years of NIH R01 grants, in that a proposal's scoring percentile explained some of the variation in the quantity of publications and citations it yielded. Yet follow up work by Fang and co-authors (2016) on the same dataset found that scores were only able to predict performance at the top percentiles and not for the majority of grants. Lauer and co-workers (2015) found no associations between percentile score and R01 grant productivity at one of the institutes at NIH, when accounting for grant amount. Kaplan et al. (2008) called for a greater number of reviewers to provide greater "statistical precision" in determining a proposal's value.

Most importantly in the context of our study, a number of scholars have put forth the critique that peer review is biased against riskier, more novel research (Braben, 2004; Chubin and Hackett, 1990; Linton, 2016; Luukkonen, 2012; Travis and Collins, 1991; Wessely, 1998). Boudreau et al. (2016) found that reviewers are systematically biased against more novel research proposals. Because high-impact ideas

[1] Much of the quantitative research on peer review focuses on US biomedical research support, specifically at the NIH, perhaps due to its size, longevity or willingness to make data accessible to researchers.

tend to also have high novelty (Foster et al., 2015; Uzzi et al., 2013; Wang et al., 2017), peer review also seems be poorly suited to selecting projects with the greatest impact.

Naturally, alternatives and modifications to peer review have been proposed to reduce bias against more innovative efforts. A wide variety of modified peer review systems have been proposed by scholars across a range of disciplines; suggestions range from adjustments in how scores are ranked to radically different systems of evaluation (Bollen et al., 2014; Casadevall and Fang, 2014; Cook et al., 2005; Johnson, 2008; Kaplan et al., 2008; Marsh et al., 2008; Roy, 1985). Linton (2016) argues that the range of peer review scores for a given project should be taken into consideration in order to counteract bias against novelty.

Despite abundant criticism and suggested alternatives, many people support retaining the general framework of peer review as it exists today. As the American Academy of Sciences put it in their report *Restoring the Foundation*, "no better system has been devised" (2014). In order make a strong argument for adopting any particular method of project selection, or even for preserving the status quo, a better understanding is needed of how each method impacts a research portfolio and its outcomes. The empirical evidence for or against any given system is scant; programs are reluctant to experiment with their procedures, and so reforms have been adopted based on intuition rather than controlled study (Azoulay, 2012; Lauer and Nakamura, 2015).

Reliance on an individual expert opinion to select proposals, as opposed to a more "democratic" system based on a set of review scores, has not been the subject of any empirical studies, to our knowledge. The canonical example of this practice is at DARPA, where scientists and engineers are hired as short-term staff members and empowered to select which proposals to fund. These program managers may seek external opinions but are not bound to act on them. Individual discretion is endorsed by those who note its use at DARPA to support novel ideas that would have been rejected by a peer review panel (Cook-Deegan, 1996; "In Defence of DARPA," 2003). However, this model is not always popular; programs that do not fund the highest-scoring projects may experience pushback and concerns over transparency and fairness (Van Noorden, 2015).

Our study of ARPA-E adds quantitative evidence to the discussion of empowering program staff to make research funding decisions. ARPA-E solicits multiple external reviews for each proposal, but they ultimately rely on the individual discretion of program director (PD) to choose which projects to fund within a technical program. This practice allows us to compare the decisions made by these staff members to the alternative decisions that could have been made based solely on the external peer reviews.

2.2. ARPA-E

In its 2007 report *Rising Above the Gathering Storm*, a committee of the National Academies called for the creation of a DARPA-like agency within the Department of Energy (DOE) to pursue transformational innovation in energy technology (National Academies, 2007). ARPA-E was tasked with "identifying and promoting revolutionary advances in fundamental science; translating scientific discoveries and cutting-edge inventions into technological innovations; and accelerating transformational technological advances in areas that industry itself is not likely to undertake because of technical and financial uncertainty" (110th Congress, 2007). To accomplish this, ARPA-E internalized the practices of active program management similar to those utilized at DARPA.

ARPA-E was established by the America COMPETES Act and first funded through the American Recovery and Reinvestment Act in 2009. Its statutory goal is to advance energy technology that reduces greenhouse gas emissions, reduces energy imports and improves energy efficiency of the US economy. ARPA-E is expected to "overcome the long-term and high-risk technological barriers in the development of energy technologies" (110th Congress 2007, sec. 5012).

As DARPA does for the US military, ARPA-E designs technical programs around specific technical challenges that could result in a transformational impact on the US energy system. The first solicitation from ARPA-E stated its intention to fund transformational "high-risk concepts with potentially high-payoff" (ARPA-E, 2009).[2] Uncertainty is therefore an inherent feature of ARPA-E's operations, due to the uncertain nature of transformational innovation compared to research that pursues incremental advances to existing technology.

One of the defining features of ARPA-E, like DARPA before it, is the autonomy granted to the PDs. A recent assessment of ARPA-E by the National Academies found that the independence of PDs is essential to the success of ARPA-E (National Academies, 2017). PDs are allotted significant discretion in selecting projects for funding, in addition to defining their own program within which to solicit projects and managing the projects after selection. It is not difficult to imagine that this autonomy in selection could have dramatic effects on the portfolio of projects that receive funding.

[2] "High-risk" here should be distinguished from scientifically unsound or unfeasible. ARPA-E solicitations state consistently that, "The proposed work may be high risk, but must be feasible."

Upon hiring, ARPA-E PDs design a program with specific technical targets, and the agency solicits research proposals through a Funding Opportunity Announcement (FOA); for an early example, see the Batteries for Electrical Energy Storage in Transportation (BEEST) FOA issued in 2010.[3] The PD then oversees the merit review process,[4] which begins with the submission of concept papers, brief summaries of proposed research ideas. ARPA-E solicits reviews of concept papers from a variety of external experts, including university-, industry-, and government-affiliated researchers. A subset of applicants is then encouraged to submit a full proposal. Full proposals include a detailed account of the research effort, milestones, timeline and budget for the proposed project.

Each full proposal is reviewed by another set of external reviewers, who provide numerical scores and comments across various attributes. Applicants are then given the chance to briefly reply to these review comments. At the end of the review process, the PD submits a recommendation to the Director of ARPA-E of which proposals to select, based on their own review of the application, the content of the external reviews, and the replies received from the applicant.[5] Proposals are then formally selected by the Director for negotiation to become a funded project, though our interviews with ARPA-E staff indicate that the vast majority of selection decisions follow the recommendation of the PD.

The implementation of individual discretion at ARPA-E is made possible by its unique organizational structure within DOE. Like DARPA's program managers, ARPA-E PDs are technical experts hired for 3-year rotations, with no expectation or possibility of career-long service to the agency and no career incentive to develop a track record by making safe choices.[6] Instead, PDs are empowered to make bold choices in pursuit of the agency's mission to accelerate transformational change in energy technology.

---

[3] The BEEST program aimed to develop "advanced battery chemistries, architectures, and manufacturing processes with the potential to provide EV [electric vehicle] battery system level energy densities exceeding 200 Wh/kg (mass density) and 300 Wh/liter (volumetric density) at system level costs of $250/kWh or below" (ARPA-E, 2010). According to the FOA, the typical cost of a lithium-ion battery system at the time was $800-$1200/kWh. Lowering the upfront cost of battery systems would open up a larger market for EVs and lead to cost savings, reduced oil imports, and reduced carbon emissions from an increasingly clean electricity supply. The BEEST program was allocated $35 million, and it funded 10 research teams from around the US including companies, universities, and national labs.

[4] Each FOA at ARPA-E is accompanied by a Merit Review Plan, which is executed by a Merit Review Board chaired by the program director that crafted the program. Our summary of the proposal and selection process is based on an example Merit Review Plan provided by ARPA-E, supplemented with information from our interviews with ARPA-E staff.

[5] Exceptions to this practice are made when the PD has a conflict of interest for a particular proposal. In this case, an alternate PD coordinates the proposal's review and manages the project if the proposal is selected.

[6] The original authorizing act for ARPA-E specifies a 3-year renewable term and the authority of the Director to

It is clear from public documents that ARPA-E's PDs are nominally empowered to use their professional judgment to allocate funding, yet the existence of this policy does not guarantee that they use this autonomy to deviate from the judgment of the external reviewers. PDs could still choose to fund only those projects that are well-liked by reviewers. Because individual discretion as a method of allocating research funding is poorly studied, it remains an open question of how these decisions are made in practice.

3. Data

To explain how individual discretion was implemented in practice in the early years of ARPA-E's existence, we compiled datasets of all proposals and funded projects in ARPA-E's funding history during two on-site visits to ARPA-E. We supplemented these datasets with intellectual property and market engagement outcomes (collected by ARPA-E), publication outcomes (collected by the authors from Web of Science), and founding year for companies (collected by the authors from public information). These datasets were merged and then scrubbed of identifying information before being removed from ARPA-E premises, in order to protect the confidentiality of the applicants.

3.1. Proposals

Our dataset of proposals contains all review scores for full proposals submitted to ARPA-E through Dec. 31, 2015. For most FOAs, reviewers rated an application on each of the following four criteria using a five-point scale, with 5 being the highest possible score:[7]

1. Impact on ARPA-E Mission Area
2. Overall Scientific and Technical Merit
3. Qualifications, Experience and Capabilities
4. Sound Management Plan[8]

hire personnel "without regard to the civil service laws" (110th Congress, 2007).

[7] Review questions for CHARGES and IDEAS did not fit this format, and so we exclude review data from those programs. We also exclude proposals for the CONNECT program, because these are for outreach projects rather than research and development.

[8] Before 2014, the ratings for "Sound Management Plan" were either "Yes" or "No". We coded these as 5 and 1 respectively.

We use the weights stated in the FOA for each component (Impact, Merit, Qualifications, and Management) to calculate an overall score for each proposal-reviewer pair.[9] One obvious shortcoming of our proposal review data is that ARPA-E's funding decisions may take into account the additional information provided in the applicant's replies to reviewer comments.

For the purpose of understanding decision-making by an individual PD, we exclude projects from "open" (non-targeted) programs, for which decision-making around project selection involved multiple PDs. Proposals in "open" programs span a wide range of technology types and are not directly compared to each other. The resulting dataset contains 1,216 proposals. Of these, 43 proposals have scores from only one external reviewer, so these are excluded from any analysis of standard deviation around mean score. 90% of proposals received 2, 3, or 4 reviews, with an average of 3 reviews. 31% of proposals in our dataset were selected to negotiate a funding award.

Table 1: Descriptive Statistics for Dataset of ARPA-E Proposals

Variable	N	Mean	S.D.	Min.	Max.
Selected	1216	0.31	0.46	0	1
Budget requested (million USD)	1216	2.81	1.86	0.14	10.00
Number of reviews	1216	3	0.92	1	7
Mean categorical scores					
Impact	1216	3.2	0.74	1.0	5.0
Merit	1216	3.1	0.75	1.0	5.0
Qualifications	1216	3.6	0.76	1.0	5.0
Management	1216	3.8	1.0	1.0	5.0
Weighted overall scores					
Mean	1216	3.4	0.69	1.0	4.9
Standard deviation	1173	0.74	0.46	0.0	2.6
Median	1216	3.5	0.76	1.0	4.9
Minimum	1216	2.7	0.92	1.0	4.9
Maximum	1216	4.0	0.71	1.0	5.0

Note: Sample is the set of ARPA-E proposals submitted 2009-2015 to targeted research programs with overall weighted review scores in the format described above.

We also look in detail at the words used by reviewers in their comments on the proposals. We choose a set of adjectives that are salient to risk tolerance, and we count the instances of each word in any of the comments within each review of a given proposal. We then subtract the instances of "not [word]" (e.g. "not novel"), leaving a simplified count of affirmative uses. For each review, we create a binary variable

[9] The most common weighting scheme was 30% each for Impact, Merit, and Qualifications, and 10% for Management. Early FOAs for Electrofuels, BEEST, and IMPACCT made no statements on category weighting, however, later programs in 2010 stated that the categories are of "equal weight," so we assigned 25% weight to each category in those FOAs.

for whether the review used a given word (W). For each proposal, we then create a continuous variable for percent of reviews that used that word (X).

Table 2: Word Occurrence in External Reviews of ARPA-E Proposals

Word (N = 3561)	Frequency of affirmative mention in review dataset (mean of W)	T-test for weighted overall score (S) [S|W=1 – S|W=0]
Positive		
innovative	23%	7.72
unique	16%	5.04
risky	4%	4.93
ambitious	5%	4.38
creative	2%	4.30
new	28%	3.23
Neutral		
unknown	3%	0.44
uncertain	4%	0.24
novel	21%	0.19
premature	1%	-0.49
Negative		
difficult	18%	-1.96
unproven	1%	-2.06
original	2%	-2.22
unrealistic	2%	-2.86
impossible	2%	-4.26

Note: Sample is the set of unique reviews of ARPA-E proposals (described above) that included comments. Two-tailed t-test with unequal variance. "Positive" defined as $p<0.10$ and $t>0$; "negative" defined as $p<0.10$ and $t<0$. Words shown here appear in the reviews of more than 1% of ARPA-E proposals.

In Table 2, we group the reviews that contain an affirmative mention of a given word (W=1) and compare their overall scores against those that do not (W=0). The sign and significance of the differences in score let us identify a few words as being significantly positive for reviewers on average, such as *innovative*, *unique*, *risky*, or *ambitious*, and some that are significantly negative, such as *unrealistic* or *impossible*.

3.2. Projects

After the applicant and ARPA-E complete negotiations on milestones, objectives, and budget, selected proposals become projects. Many ARPA-E projects are executed as partnerships between multiple organizations; for simplicity, we categorize projects by the organization type of the lead recipient. We

separate private company awardees into two categories: startups (founded no more than 5 years prior to the project start date) and established firms.[10]

We exclude projects that were still in progress in 2016 by limiting our dataset to those that ended on or before Dec. 31, 2015. As such, the latest start date for a project included in our dataset is June 2014. We also limit our dataset to only the proposals with scoring data in targeted programs, rather than "open" programs that span all areas of energy technology. The final dataset contains 165 funded projects, totaling $393 million of funding from ARPA-E.

We create an indicator variable for whether a project was selected despite a low score, and we call these projects "promoted," in the sense that they would not have been funded under a traditional peer review system. Our primary method for identifying "promoted" projects is to create a hypothetical score cutoff for each program based on the number of projects selected. We take the number of proposals selected under a given FOA to be N, and then place the cutoff at the Nth highest mean overall score. Proposals selected from scores below this cutoff are considered "promoted." This process is then repeated for rankings based on minimum overall score and maximum overall score. Because the size of a given program is limited by its budget rather than by an arbitrary number of projects, we also test alternative versions of the score cutoff based on the budget for a program rather than the number of projects selected.[11]

To address the short-term impact of ARPA-E's project selection practice, we need quantitative indicators of research progress. We use publications, patents and market engagement metrics as the outcomes of interest for ARPA-E projects, while acknowledging that these are highly imperfect indicators of value for a research project. Given the time lag on these metrics and the fact that our study period is only 5 years long, we are only able to capture an early glimpse at the productivity of the funded projects.

---

[10] The primary mechanism for ARPA-E funding is a cooperative agreement. When a national lab participates in a project, whether or not it is the lead recipient, it is funded separately through a contract mechanism. Additionally, some non-lead members of a project team may have a separate award issued to their organization during the course of the project. In these cases, we combine the data for multiple awards into a single project. As a result, our unit of analysis is a cohesive technical effort by a team of researchers.

[11] In the budget-based method, we tally the cumulative proposed budgets of the proposals to a given FOA, starting from the highest mean overall score, until this cumulative budget reaches the total budget listed in the FOA. For nearly every program, this method produced a higher score cutoff than the one based on number of projects; we focus our analysis on the projects-based metric to obtain a conservative estimate of the number of "promoted" projects.

Publication data were collected for each award through Dec. 31, 2015. We collected these data by searching Web of Science for the award or work authorization numbers for the funded ARPA-E projects in our dataset. Some publications are flagged as "highly cited" if they exceed the top percentile of citations for papers published in the same year and journal subject category.

Awardees are required as part of their cooperative agreement to acknowledge ARPA-E support in any patents and also to report intellectual property to DOE. ARPA-E has, in collaboration with the DOE General Counsel's office, collected data on invention disclosures, patent applications, and patents issued as a result of each project. We obtained these data from ARPA-E on inventive outcomes for each award through Dec. 31, 2015.

ARPA-E also tracks the progress of awardees in market engagement. Each spring, to coincide with their annual summit, ARPA-E publishes a list of projects that have received (i) follow-on private funding, (ii) those that have additional government partnerships and (iii) those that have formed companies.[12] We separately obtained from ARPA-E a list of awards that have led to (iv) initial public offerings (IPOs), (v) acquisitions, or (vi) commercial products. All of these outputs are those that the awardee reports as being directly attributable to ARPA-E support. We also obtained the dollar amounts of private funding deals, when these were reported to ARPA-E. Our market engagement data are through February 2016.

We created two aggregated metrics which combine the three categories of external outputs that we measure: publications, inventions and market engagement. First, we measure whether a project produced at least one external sign of progress: a publication, a patent application, *or* some form of market engagement (among the six types of market engagement measured). Second, we measure whether a project received all three of the key metrics: a publication, a patent application, *and* some form of market engagement.

[12] "Company formation" for our purposes includes startup company awardees for which the ARPA-E award was their first funding.

Table 3: Descriptive Statistics for Dataset of ARPA-E Projects

Variable	Mean	S.D.	Min.	Max.
Initial project length (years)	2.22	0.76	0.42	3.04
Final project length (years)	2.72	1.02	0.38	5.00
Initial award amount (million USD)	2.14	1.41	0.20	6.00
Final award amount (million USD)	2.38	1.53	0.20	6.67
"Promoted"				
Low mean score	0.55	0.50	0	1
Low min. score	0.52	0.50	0	1
Low max. score	0.47	0.50	0	1
All around (3/3)	0.28	0.45	0	1
Not at all (0/3)	0.27	0.45	0	1
External outputs				
At least 1 publication	0.45	0.50	0	1
At Least 1 patent application	0.42	0.50	0	1
Market engagement	0.32	0.47	0	1
Any external outputs (>0 of 3)	0.75	0.44	0	1
All external outputs (3 of 3)	0.08	0.28	0	1

Note: Sample is the set of ARPA-E projects completed 2009-2015 (N = 165) within targeted research programs with overall weighted review scores in the format described in Section 3.1. The "promoted" variable marks whether a proposal was selected despite a score below a hypothetical cutoff, calculated based on the number of projects funded in a given technical program. Project outputs are measured through Dec. 31, 2015.

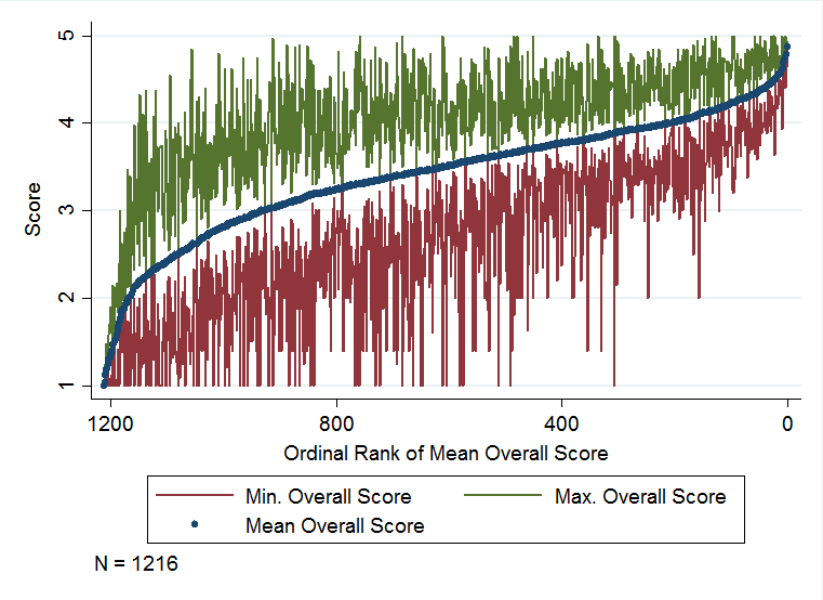
In addition to the proposal and project data described above, we interviewed nine current and former ARPA-E technical staff members. Qualitative information provided in those interviews informs our analyses below.

4. Results

In this section, we describe the project selection decisions made by ARPA-E PDs and establish that these decisions were different than they would have been in a counterfactual peer-reviewed program. We then use our data to seek explanations for the how these decisions are made. We conclude with an early effort to evaluate how individual discretion influences research outputs across the portfolio.

We observe wide variation in review scores across the entire set of ARPA-E proposals (Figure 1). The mean score varies across the full scale from 1 to 5, and the standard deviation of scores around the mean is as high as 2.6 for some proposals (Table 1).

Figure 1: Scores of ARPA-E Proposals Ranked by Mean Score



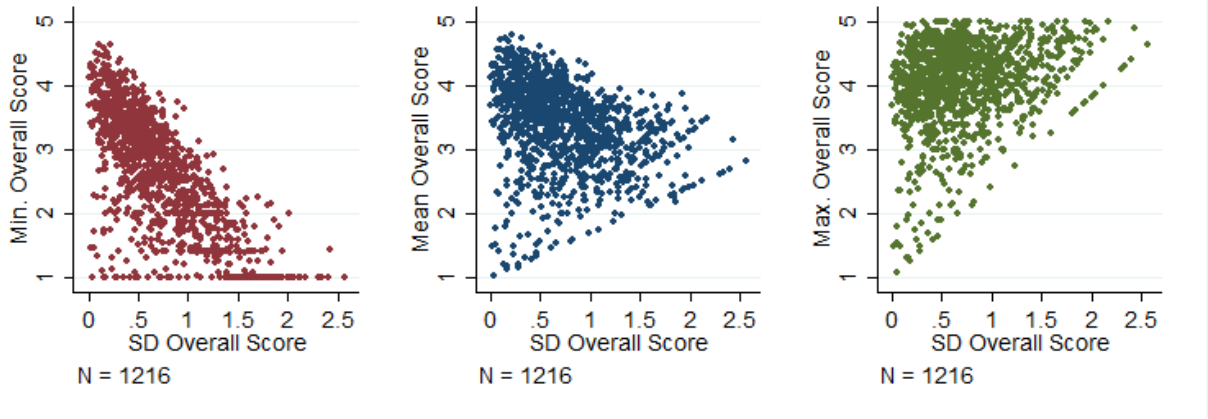
Note: Proposals to ARPA-E ranked in order of mean score, with minimum and maximum scores also plotted.

As a stylized construct to represent traditional peer review, we consider three scoring methods that could be used to rank proposals:

1. *Minimum score rank*: Select for proposals with no detractors, by only considering the lowest score received by each
2. *Mean score rank*: Weigh every score equally in determining a proposal's quality
3. *Maximum score rank*: Select for proposals with champions, by only considering the highest score received by each

Importantly, these three score-ranking methods have different implications for the extent of uncertainty in the funded portfolio. The extent of possible disagreement decreases as the mean score approaches its upper limit of 5.0; this concept is illustrated in Figure 2. Methods 1 and 2 (ranking on minimum or mean scores) mechanically limit the amount of disagreement that will be tolerated. Selecting the highest maximum scores, on the other hand, places no restrictions on the extent of disagreement. A proposal may receive a maximum score of 5.0, regardless of whether other reviewers scored it as a 1.0.

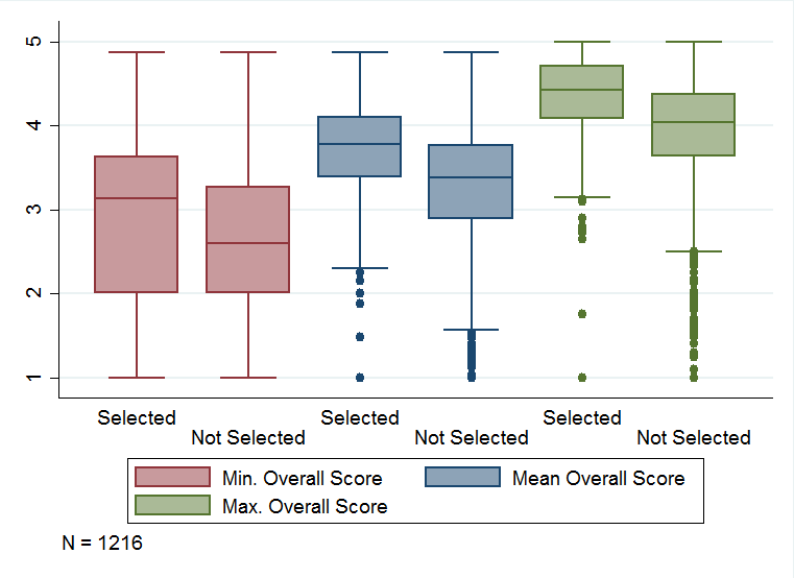
Figure 2: Scoring Statistics for ARPA-E Proposals



Note: Each plot depicts an element of the score distribution for proposals submitted to ARPA-E vs. the standard deviation of those scores.

Do ARPA-E's selection decisions resemble any of these three score-ranking methods? We compare the overall scores for funded and unfunded proposals and find that none of the three is the sole deciding factor for PDs when selecting proposals (Figure 3). All three measures of the score distribution (min., mean and max.) are higher for funded proposals, and yet there is significant overlap of scores between funded and unfunded proposals. Some projects were selected despite very low scores, and some were not selected despite very high scores.

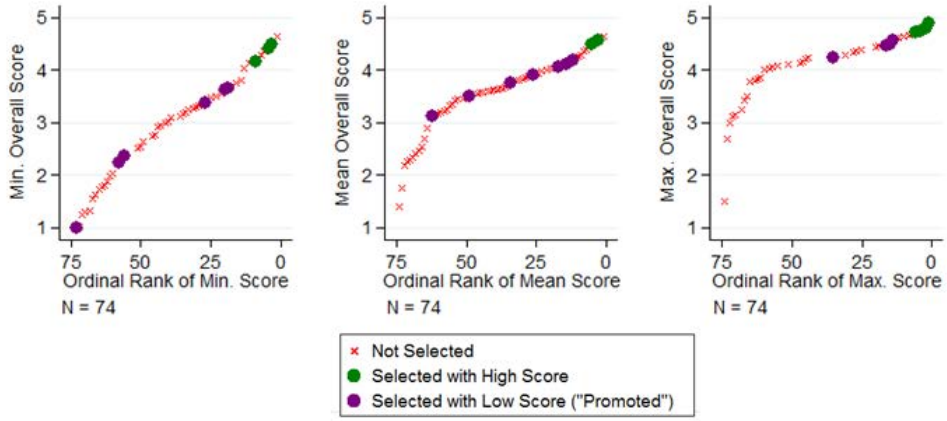
Figure 3: Box Plots of Scoring Statistics for Unfunded and Funded ARPA-E Proposals



Note: Modified box plot depicts percentiles of score distributions for funded and unfunded ARPA-E proposals. Outside values (< 25th percentile – 1.5 · interquartile range) are plotted as points.

The data depicted in Figure 3 represent proposals aggregated across 33 different technical programs. The conditions for program directors' decision-making varied significantly between programs—for example, in the funding available, or the number of proposals submitted—and so in the remaining results, we focus on selection at the level of the technical program. Comparing the scores of funded and unfunded proposals in a single technical program at ARPA-E (Figure 4), we see that proposals were selected for funding from across the full range of scores, including many that were "promoted."

Figure 4: Proposals to the BEEST Program



Note: Scores for proposals to the ARPA-E Batteries for Electrical Energy Storage in Transportation (BEEST) program, shown in order of three ranking criteria. Of 74 proposals, 9 were selected for funding. "Promoted" proposals were selected despite a score below a hypothetical cutoff, which was the score of the 9th highest scoring proposal.

Having shown that ARPA-E PDs use significant individual discretion to select proposals, we move on to explore how these decisions correlate with external review scores.

4.1. Can review score characteristics explain selection decisions?

The use of discretion does not imply that selection decisions are fully unrelated to scores. After all, ARPA-E program directors are technical experts and members of the same research community from which external reviewers are sourced; it would be a surprise if their judgments had no relationship whatsoever to those of the reviewers. Indeed, we see evidence in the aggregate data that scores do correlate somewhat with selection decisions. Nearly half of selected projects across the entire set of ARPA-E proposals are "promoted" from a low score: 55% "promoted" by mean score, 52% "promoted" by minimum score, and 47% "promoted" by maximum score (Table 3). If there were a uniform probability of selection across scores, then the "promoted" label would be even more common—closer to the overall rejection rate for full proposals, which is 69%.

Our interviews with ARPA-E staff indicate that there is a culture of risk-taking at the agency, which pushes PDs to select some proposals that are not uniformly liked by external reviewers. This risk-taking could manifest in our dataset in several ways, such as a preference for a wide range of scores, or a preference for low minimum scores.

To understand how scores might influence PD decisions, within the constraints of this empirical setting, we conduct regression analysis to model the predictive power of scores on selection decisions, using a linear probability model:

$$Y_i = \alpha_0 + \alpha_1 Score_i + \varphi_i + \varepsilon_i$$

Y_i is the binary outcome variable for whether proposal *i* was selected; $Score_i$ is the scoring element of interest for proposal *i*, e.g. mean overall review score; φ_i is a fixed effect for the technical program. Our choice of a linear probability model is based on the ease of interpretation for these results, shown in Table 4. Similar results using a logit model are shown in the Appendix (Table A1).

Our three simple score heuristics (mean, minimum, and maximum score) are all individually predictive of selection at ARPA-E. Of the three, mean score has the strongest correlation; there is a 19% greater probability of selection for each additional point in the mean overall score. Yet the R^2 value is relatively low (0.13), indicating that only a small portion of variation in selection is explained by the mean score. Minimum score has both the weakest correlation (8% increased probability of selection per additional point) and the least explanatory value ($R^2 = 0.08$).

Table 4: Predicting Selection by Review Score Distribution

Dependent Variable: ARPA-E Selected Proposal for Funding	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Mean Overall Score	0.194***				0.251***		
	(0.043)				(0.029)		
Min. Overall Score		0.078*				0.026	-0.033
		(0.044)				(0.040)	(0.042)
Max. Overall Score			0.174***			0.161***	0.056*
			(0.029)			(0.021)	(0.031)
SD Overall Score				0.060	0.137**		
				(0.069)	(0.066)		
Med. Overall Score							0.164***
							(0.033)
Program F.E.	Y	Y	Y	Y	Y	Y	Y
N	1216	1216	1216	1173	1173	1216	1216
R^2	0.131	0.080	0.123	0.064	0.164	0.125	0.143

Notes: Standard errors in parentheses. All regressions are OLS with robust standard error, clustered by technical program.
* p < 0.10, ** p < 0.05, *** p < 0.01

Several additional relationships between score and selection are explored in the Appendix. The overall score is broken down into its components, showing that scores for Merit and Impact are predictive of

selection, while scores for Qualifications and Management have little to no statistical relationship to selection (Table A2). In Table A3 and Table A4, we reproduce the analyses from Table 4 for Merit and Impact scores separately, with nearly identical results to the trends on overall score.

In order qualify the strength of the correlations between score and selection, we compare them to the relationship between score and hypothetical selection under each of three systematic selection methods (Table A5). We find that the linear coefficients predicting selection for both mean score and maximum score (0.194 and 0.174, respectively, in Table 4) are less than half what they would be if ARPA-E selected projects by ranking those scores (0.468 and 0.415 in Table A5). For minimum score, the association is even weaker: the coefficient predicting actual selection is five times smaller than the coefficient predicting a high minimum score.

We also estimate the predictive power of reviewer disagreement on selection at ARPA-E. The standard deviation of overall scores for a proposal does not significantly correlate on its own with selection for a given program (Model 4), but it has a positive and significant coefficient when controlling for the mean overall score (Model 5). In other words, ARPA-E PDs tend to fund proposals on which reviewers disagree, given the same mean overall score.

The tendency of ARPA-E PDs to select projects with a wide spread of scores is not necessarily symmetric around the mean score. In fact, when minimum and maximum score are both accounted for, the coefficient on minimum score disappears. This suggests that ARPA-E PDs prefer proposals that were highly rated by at least one reviewer, but they are not deterred by the presence of a low rating. This trend persists when median score is included (Model 7 in Table 4). ARPA-E PDs tend to agree with the bulk of reviewers, and they also tend to agree with scores in the upper tail of the distribution. They use their discretion to surface proposals that have at least one champion, regardless of whether there are any detractors.

The number of external review scores recorded for proposals in our primary dataset ranges from 1 to 7. Standard deviation is of course a less reliable measure of reviewer disagreement for a very small set of reviews, so in the Appendix, we exclude the proposals with less than 3 reviews and repeat the analyses above. The findings above are robust, except that the coefficient on standard deviation in Model 5 loses significance (Table A6). Yet the coefficient on maximum score in Model 7 gains in both size and significance, confirming that ARPA-E PDs tend to select projects specifically with *upside potential*, indicated by the presence of a high rating by at least one reviewer.

Another way to measure the uncertainty associated with "promoted" proposals is to compare them to those that were rejected, despite high scores—we call this category of proposals "demoted." Compared to a counterfactual program that select projects based solely on scores, the effect of individual discretion is to replace "demoted" projects with "promoted" projects. In the Appendix, we show that the spread of scores for "promoted" proposals is greater than for "demoted" proposals on the basis of mean score (Table A7), even beyond the mechanical association of mean score and standard deviation. These results support our finding that the use of individual discretion serves to allow uncertainty in the ARPA-E portfolio.

The regressions above implicitly give equal weight to each reviewer's opinion, by giving equal consideration to each score in the score distribution for a proposal. Unfortunately, we are not able to investigate whether PDs give different weights to individual reviewers' scores, because each reviewer typically only reviews a small sample of proposals; the median number of reviews is 6 among the set of 553 reviewers in our dataset, and 79% of reviewers only review proposals within a single technical program. However, interviews with ARPA-E staff indicate that the PD's knowledge of the reviewer can at times determine the weight he or she gives that review. Such weighting could include discounting a reviewer's opinion, when the topic of the proposal is far from their area of expertise, or observing a reviewer's tendency to give especially low scores.

4.2. Can review comments explain project selection?

Next, we look beyond the numeric scores and ask whether the information contained in the review comments offers some explanation of ARPA-E PDs selection decisions. In our interviews, multiple ARPA-E PDs described ignoring the review scores and instead considering the written content of the reviews. In an attempt to capture this phenomenon, we measure the correlation of selection with certain descriptive words in the review comments.

Table 5: Predicting Selection by Word Occurrence

Word (N = 1209)	Coefficient on X predicting project selection	
	Alone	With mean overall score
Positive		
innovative	0.155**	0.083
	(0.065)	(0.060)
unique	0.104	0.013
	(0.070)	(0.073)
risky	0.195	0.077
	(0.117)	(0.123)
ambitious	0.175	0.142
	(0.132)	(0.131)
creative	0.444***	0.382***
	(0.156)	(0.136)
new	0.010	-0.003
	(0.045)	(0.043)
Neutral		
unknown	-0.091	-0.045
	(0.112)	(0.113)
uncertain	-0.123	-0.135
	(0.091)	(0.094)
novel	0.033	0.031
	(0.051)	(0.050)
premature	0.195	0.227
	(0.190)	(0.186)
Negative		
difficult	-0.110*	-0.088
	(0.062)	(0.059)
unproven	-0.173	-0.107
	(0.113)	(0.120)
original	-0.109	-0.079
	(0.123)	(0.134)
unrealistic	0.062	0.162
	(0.176)	(0.173)
impossible	-0.311*	-0.045
	(0.181)	(0.170)

Note: "Positive" defined as $p<0.10$ and $t>0$; "negative" defined as $p<0.10$ and $t<0$ in two-tailed t-test with unequal variance from Table 2. Sample is the set of ARPA-E proposals (described in section 3.1), excluding seven proposals that had no review comments. Columns 2 and 3 are regression coefficients for OLS models of project selection (with program fixed effects) based on the percent of reviews featuring that word as a description, either alone (Column 3) or controlling for the mean overall score of that proposal (Column 4).

Relatively few of these descriptive words predicted selection: reviewers' use of *innovative* and *creative* increases with the probability of selection, and use of *difficult* and *impossible* decreases with probability of selection. Controlling for mean score, however, eliminates three out of four of these trends, such that ARPA-E PDs do not appear to take any additional information from those descriptions beyond the score.

Use of the word *creative*, however, stands out as having predictive power beyond the mean overall review score. When a reviewer describes some element of the proposal as *creative*, ARPA-E PDs are on average significantly more likely to fund that proposal, even over other proposals in the same program with the same average score.

4.3. What else might explain project selection?

There are alternative inputs for project selection that cannot be tested quantitatively using the data in this study. Supporting evidence for two additional explanatory factors emerge from the qualitative data collected through interviews with ARPA-E staff.

1. Selection could be based on the need for a diverse portfolio of technical approaches

Within each FOA for a program at ARPA-E, there are often several possible approaches outlined to reaching the technical targets; the BEEST program, for example, funded projects aimed at new anode materials, new manufacturing processes, and non-lithium battery designs. Our interviews with PDs and ARPA-E leadership indicate that diversity of approaches is desirable, such that PDs may choose to fund a portfolio that includes multiple approaches, making direct comparisons only within smaller groups of similar proposals. PDs are encouraged to construct a portfolio of different types of technology, in order to maximize the chance that one or more projects will be able to achieve the targets set out in the FOA. The value of each project to a diverse portfolio may not be obvious to external reviewers, who only see a small sample of proposals, but can be taken into account by a PD with a holistic view of the program and the goals of the agency in mind.

2. Selection could be based on idiosyncratic PD preferences

Because of the autonomy given to each PD, their decisions may be dominated by their own individual preferences, rather than any discernible trends over the entire agency. There could be variation between individual PDs in their tolerance for risk and treatment of low scores. Each PD could also apply their own judgment to specific research activities, based on their professional experience. After all, ARPA-E PDs often have the same career profile as the “peers” from whom they solicit external reviews.

Although we cannot quantify these idiosyncratic effects without much larger samples of decisions made by each PD across multiple technical programs, we find evidence in our interviews that ARPA-E PDs differ somewhat in their attitudes regarding project selection. For example, some PDs report constructing

a portfolio that is diverse along the dimension of risk, so that each program has a mix of safe and risky projects, while other PDs report a more uniform preference for high-risk projects.

4.4. Short-term impact of selection method

Next, we look for the effect of PD selection decisions on the performance of ARPA-E's research portfolio. In particular, we are interested in the performance of "promoted" projects, which would have been rejected if not for individual discretion. Unfortunately, this analysis is limited, because we cannot observe outputs from "demoted" projects, those that would have been *accepted* if not for individual discretion. Additionally, it is important to note that even the longest-running projects in our dataset began only six years before the end of our study period; it is too early to observe the full extent of research outputs from the ARPA-E portfolio. Nonetheless, we endeavor to see how the additional uncertainty baked into the portfolio through "promoted" projects affects their research outputs, compared to those projects that were uncontroversial.

Modeling the probability of research outputs requires that we test the inclusion of several control variables, as there are inherent features of a project that can impact the rate of publishing, patenting and/or market activity. Outputs may be associated with both the organization type (university, established firm, startup, non-profit or National Lab) and project funding amount. Here we control for the initially negotiated project budget, in order to compare projects that were prospectively similar at the outset. The final funding amount is endogenous, as many of the award budgets were adjusted mid-project, and these adjustments likely related to project performance.

We ask whether the projects that were "promoted" by individual discretion of ARPA-E PDs were more or less productive (in terms of publications, patent applications, or market engagement metrics) compared to the projects with high review scores, using the following regression model:

$$Y_i = \alpha_0 + \alpha_1 "Promoted"_i + \alpha_2 ln(initial\ funding\ amount_i) + \varphi_i + \delta_i + \varepsilon_i$$

Y_i in this case is the binary outcome variable for whether project *i* resulted in a given output measure; $"Promoted"_i$ is a binary indicator for whether the proposal for project *i* scored below a hypothetical score cutoff; φ_i is a fixed effect for the technical program; δ_i is the fixed effect for the type of organization leading the project. In Table 6, we test several models for the relationship between "promoted" (on the basis of mean overall scores) and one particular output: whether or not a project produced a publication. No association is found, regardless of the control variable structure.

Table 6: Control Variables for Publication Output

Dependent Variable: At Least 1 Publication from ARPA-E Project	(1)	(2)	(3)	(4)	(5)
"Promoted"	0.051	0.067	-0.074	-0.056	-0.055
(low mean score)	(0.089)	(0.094)	(0.074)	(0.079)	(0.076)
Program F.E.		Y	Y	Y	Y
Org. Type F.E.			Y	Y	Y
Initial Award Amount				Y	
Log of Initial Award Amount					Y
N	165	165	165	165	165
R^2	0.003	0.162	0.362	0.367	0.367

Note: Standard errors in parentheses. All regressions are OLS with robust standard error, clustered by technical program.
* p < 0.10, ** p < 0.05, *** p < 0.01

Next, we test all three "promoted" variables for an association with multiple project outputs. The estimations in Table 7 largely show no significant difference in external measures of short-term performance between low-scoring and high-scoring projects. Projects that are "promoted" from a low review score are nearly indistinguishable in terms of output from those that would have been selected even without individual discretion, within the error of our measurement. Of the 15 regressions shown in Table 7, there are two exceptions: decreased probability of a patent application and of achieving any single measure of progress from those projects that received a low minimum score. In the Appendix, we show the same 15 regressions using alternative independent variables: (i) mean, min. and max. review scores (Table A8) and (ii) an alternative calculation of the "promoted" variable based on the program budget (Table A9). Again, we find no consistent trends.

Table 7: Outputs of "Promoted" Projects

Dependent Variable:	(1) At Least 1 Publication	(2) At Least 1 Patent Application	(3) Market Engagement	(4) Any External Output	(5) All External Outputs
"Promoted"	-0.055	0.026	0.031	0.042	-0.016
(low mean score)	(0.076)	(0.116)	(0.085)	(0.071)	(0.076)
N	165	165	165	165	165
R^2	0.367	0.327	0.284	0.362	0.163
"Promoted"	-0.023	-0.140**	0.008	-0.109*	-0.039
(low min. score)	(0.086)	(0.066)	(0.075)	(0.062)	(0.053)
N	165	165	165	165	165
R^2	0.366	0.339	0.284	0.370	0.165
"Promoted"	0.074	-0.047	0.073	0.047	0.052
(low max. score)	(0.073)	(0.081)	(0.091)	(0.064)	(0.061)
N	165	165	165	165	165
R^2	0.370	0.328	0.288	0.363	0.169

Notes: Standard errors in parentheses. All regressions are OLS with robust standard error, clustered by technical program. The models include controls for the log of initial award amount, as well as a fixed effect for technical program and a fixed effect for the organization type (same as Model 5 in Table 6).
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

As a further check for differences in project performance among proposals with different scores, we create two additional variables: one for being "promoted" on the basis of all three score rankings (mean, min., and max.) and one for not being "promoted" at all—for the group of projects that would have been selected by any of the three ranking methods. The regressions in Table 8 further demonstrate that the projects selected via PD discretion have roughly equivalent performance to those that were uncontroversial (i.e. scored highly in external review).

Table 8: Outputs of "Promoted" All Around vs. Not at All "Promoted" Projects

	(1) At Least 1 Publication	(2) At Least 1 Patent Application	(3) Market Engagement	(4) Any External Output	(5) All External Outputs
"Promoted" all around (low mean, min., *and* max. score)	0.066 (0.092)	-0.220 (0.144)	0.040 (0.106)	-0.032 (0.075)	0.010 (0.070)
Not "promoted" by any measure	0.014 (0.082)	-0.086 (0.083)	-0.086 (0.056)	-0.012 (0.085)	-0.037 (0.056)
N	165	165	165	165	165
R^2	0.368	0.351	0.291	0.361	0.165

Notes: Standard errors in parentheses. All regressions are OLS with robust standard error, clustered by technical program. The models include controls for the log of initial award amount, as well as a fixed effect for technical program and a fixed effect for the organization type. The base category is projects that were "promoted" by one or two measures, but not all three.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Beyond the five metrics above, we also consider several other short-term research output metrics: volume of publications, publications that receive relatively high numbers of citations, patents issued rather than simply applied for, and the amount of private funding obtained. Regressions of these outputs vs. a project's "promoted" status (shown in Table A10 in the Appendix) also point to an equivalence between projects that were well-liked by external reviewers and those that were not.

5. Discussion

Our results show that the portfolio of ARPA-E projects selected under individual discretion differs significantly from the portfolio that would be selected by ranking of peer review scores. Approximately half of the selection decisions made by ARPA-E PDs diverge from the aggregated opinions of external reviewers, whether that opinion is measured by the mean score or by either extreme of the distribution. In terms of explanatory variables, we find that proposals are more likely to be selected if reviewers disagree on the quality of the proposal, particularly if the proposal has at least one champion, and that proposals are more likely to be selected if reviewers describe them as *creative*. We also find that "promoted" projects perform equally well on average to non-"promoted" projects on short-term metrics. Taken together, these results show that ARPA-E PDs use their autonomy and that the content of the agency's portfolio is different as a result, in that it is enriched in technically risky but creative ideas, without any tradeoff in short-term research outputs.

An important caveat is that our findings describe correlations, rather than a causal effect of any particular feature of a proposal on its chances of being selected. Program directors may be influenced causally by scores or comments in external reviews, or they may base their decisions entirely on unobserved variables that happen to correlate with external reviews. Rather than attempting to describe the mindset of each program director when making selections, we describe the *ex ante* qualities of selected proposals and, by extension, ARPA-E's research portfolio as a whole.

We focus on each proposal's overall score, which is a composite of a reviewers' assessment of the four review criteria (Impact, Merit, Qualifications, and Management). The selection trends related to overall score appear to be driven equally by the scores for Merit and Impact. Interestingly, these two scoring categories are most closely analogous to two sources of uncertainty related to research activities: (i) technical uncertainty, i.e. the inability to predict whether a project will achieve its technical goals, and (ii) value uncertainty, i.e. the inability to predict whether, conditional on technical success, the targeted technology will bring societal benefit. ARPA-E PDs decisions to select proposals with champions serves to welcome both types of uncertainty into the portfolio, similar to other champion-based funding programs, such as the Gates Foundation's Grand Challenges Exploration programs (Grand Challenges, 2016) and some angel financing groups (Kerr et al., 2014).

Our analyses of reviewer word usage add color to our understanding of how PDs select projects, beyond whether or how they use numeric scores. In particular, the preferential selection of *creative* proposals provides an interesting contrast to the emphasis in other funding agencies on giving researchers themselves the freedom to explore different research directions (Azoulay, 2011). In the case of ARPA-E, it seems that the PD themselves are acting as agents of innovation through their use of discretion as subject matter experts.

In terms of how ARPA-E's openness to uncertainty has translated into a measurable difference in the impact of its portfolio, we are cautious in interpreting our results. Because our study design does not capture outputs from unfunded proposals, we cannot directly compare the set of funded ARPA-E projects to the project ideas that were rejected. It could be that ARPA-E projects on average perform better, worse, or the same as the projects that would have been funded using alternative selection methods—although we note that research published elsewhere shows high performance of ARPA-E projects on patenting and publication outcomes, compared to other funding sources within DOE (Goldstein and Narayanamurti, 2018) and for startup awardees in particular compared to other cleantech startups (Goldstein et al., 2020).

We recognize that publications, patent applications and market engagement are not measures of success in themselves. Rather, they are early signs of progress toward the ultimate goal of ARPA-E, which is to have a transformational impact on the US energy system. This impact will take decades to materialize as the technologies created with ARPA-E support are developed and deployed. Projects selected using individual discretion at ARPA-E may ultimately have more divergent long-term outcomes, due to a higher level of uncertainty *ex ante*. Even if many ARPA-E projects result in technical failure, the costs of those projects could be dwarfed by the returns on just a few hugely impactful projects. Indeed, DARPA's support of research in the 1950's and 1960's that led to the development of the internet and the Global Positioning System (Alexandrow, 2008; Waldrop, 2008) is often invoked to justify the public investment in DARPA over the years. Time will tell what kind of technological change ARPA-E's funding brings about in the long-term.

In designing a research program, administrators must choose strategies that best fit their organization's goals. In this paper, we show that managing project selection through individual discretion can result in a broadly different portfolio than managing through peer review. The potential for introducing uncertainty is useful for an agency like ARPA-E, which targets high-risk, high-reward research; it may be inappropriate for an agency that is not similarly mission-oriented. It is also important to note that the selection decision is only one of many elements of program design. There is likely an interplay between the mode of project selection and other program features, such as organizational structure and project management strategy. These complementarities would be a productive avenue for future research.

6. Conclusion

Like DARPA before it, ARPA-E uses individual discretion to select which research projects to fund. By also collecting external reviews of the proposed projects, ARPA-E has provided a unique opportunity to study the impact of this approach. In this paper, we provide the first quantitative study of individual discretion as a method of project selection for a research funding agency.

We show that ARPA-E program staff use their significant autonomy to construct a research portfolio that diverges from the consensus of external peer reviewers. As a result of this practice, ARPA-E was able to counter the conservatism associated with traditional peer review and fund a research portfolio encompassing more uncertainty and creativity.

The fact that program directors are nominally empowered at ARPA-E did not guarantee this outcome for the agency. Empowered program directors could have chosen instead to fund projects that aligned with

the average opinions of external reviewers. Our findings here point to an intentional choice on the part of the agency to encourage its staff to make bold choices, in line with its mission to pursue transformational energy research.

Acknowledgements

Our analysis originated as a consulting engagement with the National Academies of Science, Engineering and Medicine for a study on ARPA-E (National Academies, 2017). A. P. G. was supported by a fellowship from the Belfer Center for Science and International Affairs. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors. We are thankful for helpful discussions with Laura Diaz Anadon, Pierre Azoulay, Paul Beaton, Iain Cockburn, Gail Cohen, Jeff Furman, Daniel Kim, Josh Krieger, Gilbert Metcalf, Ramana Nanda, Venky Narayanamurti, Scott Stern, and participants in the NBER Productivity and Innovation seminar. We also thank ARPA-E staff who assisted with data collection, in particular Dave Dixon, Ron Faibish, Andy Kim and Ashley Leasure. All errors or omissions are our own.

References

110th Congress, 2007. America COMPETES Act. United States.

Alexandrow, C., 2008. The Story of GPS, DARPA: 50 Years of Bridging the Gap.

ARPA-E, 2009. OPEN 2009 Program Overview.

ARPA-E, 2010. BEEST Program Overview.

Arrow, K., 1962. Economic welfare and the allocation of resources for invention, in: The Rate and Direction of Inventive Activity: Economic and Social Factors. pp. 609–626. https://doi.org/10.1521/ijgp.2006.56.2.191

Azoulay, P., 2012. Research efficiency: Turn the scientific method on ourselves. Nature 484, 31–32. https://doi.org/10.1038/484031a

Azoulay, P., Graff Zivin, J.S., Manso, G., 2011. Incentives and Creativity: Evidence from the Academic Life Sciences. RAND J. Econ. 42, 527–554.

Baldwin, M., 2017. In referees we trust? Phys. Today 70, 44–49. https://doi.org/10.1063/PT.3.3463

Bollen, J., Crandall, D., Junk, D., 2014. From funding agencies to scientific agency. EMBO Rep. 15, 1–3. https://doi.org/10.1002/embr.201338068

Bonvillian, W.B., van Atta, R., 2011. ARPA-E and DARPA: Applying the DARPA model to energy innovation. J. Technol. Transf. 36, 469–513. https://doi.org/10.1007/s10961-011-9223-x

Boudreau, K.J., Guinan, E., Lakhani, K.R., Riedl, C., 2016. Looking Across and Looking Beyond the Knowledge Frontier: Intellectual Distance and Resource Allocation in Science. Manage. Sci. forthcoming. https://doi.org/10.1287/mnsc.2015.2285

Braben, D.W., 2004. Pioneering research: a risk worth taking. Wiley.

Casadevall, A., Fang, F.C., 2014. Taking the Powerball Approach to Funding Medical Research. Wall Str. J. A15.

Chubin, D.E., Hackett, E.J., 1990. Peerless science: peer review and U.S. science policy. State University of New York Press.

Cook-Deegan, R.M., 1996. Does NIH need a DARPA? Issues Sci. Technol. 13, 25.

Cook, W.D., Golany, B., Kress, M., Penn, M., Raviv, T., 2005. Optimal Allocation of Proposals to Reviewers to Facilitate Effective Ranking. Manage. Sci. 51, 655–661. https://doi.org/10.1287/mnsc.1040.0290

Dasgupta, P., David, P.A., 1994. Toward a new economics of science. Res. Policy 23, 487–521. https://doi.org/10.1016/0048-7333(94)01002-1

Fang, F.C., Bowen, A., Casadevall, A., 2016. NIH peer review percentile scores are poorly predictive of grant productivity. Elife 5, 1–6. https://doi.org/10.7554/eLife.13323

Foster, J.G., Rzhetsky, A., Evans, J.A., 2015. Tradition and Innovation in Scientists' Research Strategies. Am. Sociol. Rev. 80, 875–908. https://doi.org/10.1177/0003122415601618

Fuchs, E.R.H., 2010. Rethinking the role of the state in technology development: DARPA and the case for embedded network governance. Res. Policy 39, 1133–1147. https://doi.org/10.1016/j.respol.2010.07.003

Gillespie, G.W., Chubin, D.E., Kurzon, G.M., 1985. Experience with NIH Peer Review: Researchers ' Cynicism and Desire for Change. Sci. Technol. Hum. Values 10, 44–54.

Goldstein, A. P., Narayanamurti, V., 2018. Simultaneous pursuit of discovery and invention in the US Department of Energy. Res. Policy 47, 1505–1512.

Goldstein, A. P.; Doblinger, C.; Baker, E.; Anadon, L. D., 2020. Startups supported by ARPA-E were more innovative than others but an investment gap may remain. Nature Energy 5, 741–742.

Gordon, R., Poulin, B.J., 2009. Cost of the NSERC science grant peer review system exceeds the cost of giving every qualified researcher a baseline grant. Account. Res. 16, 232–233. https://doi.org/10.1080/08989620903065590

Grand Challenges, 2016. How Grand Challenges Explorations Grants Are Selected [WWW Document]. URL http://gcgh.grandchallenges.org/how-grand-challenges-explorations-grants-are-selected. Accessed 2016.

Guthrie, S., Guérin, B., Wu, H., Ismail, S., Wooding, S., 2013. Alternatives to peer review in research project funding.

In Defence of DARPA, 2003. . Nature 424, 599. https://doi.org/10.1017/epi.2014.33

Ismail, S., Farrands, A., Wooding, S., 2009. Evaluating Grant Peer Review in the Health Sciences. A review of literature.

Johnson, V.E., 2008. Statistical analysis of the National Institutes of Health peer review system. Proc. Natl. Acad. Sci. U. S. A. 105, 11076–11080. https://doi.org/10.1073/pnas.0804538105

Kaplan, D., Lacetera, N., Kaplan, C., 2008. Sample size and precision in NIH peer review. PLoS One 3, 3–5. https://doi.org/10.1371/journal.pone.0002761

Kerr, W.R., Lerner, J., Schoar, A., 2014. The consequences of entrepreneurial finance: Evidence from angel financings. Rev. Financ. Stud. 27, 20–55. https://doi.org/10.1093/rfs/hhr098

Knight, F., 1921. Risk, uncertainty and profit.

Lauer, M.S., Danthi, N.S., Kaltman, J., Wu, C., 2015. Predicting Productivity Returns on Investment. Circ. Res. 117, 239–243. https://doi.org/10.1161/CIRCRESAHA.115.306830

Lauer, M.S., Nakamura, R., 2015. Reviewing Peer Review at the NIH. N. Engl. J. Med. 373, 1893–1895. https://doi.org/10.1056/NEJMp1002530

Lee, C.J., Sugimoto, C.R., Zhang, G., Cronin, B., 2013. Bias in Peer Review. J. Am. Soc. Inf. Sci. Technol. 64, 2–17. https://doi.org/10.1002/asi

Li, D., Agha, L., 2015. Big names or big ideas: Do peer-review panels select the best science proposals? Science 348, 434–438. https://doi.org/10.1126/science.aaa0185

Linton, J.D., 2016. Improving the Peer review process: Capturing more information and enabling high-risk/high-return research. Res. Policy 45, 4–6. https://doi.org/10.1016/j.respol.2016.07.004

Luukkonen, T., 2012. Conservatism and risk-taking in peer review: Emerging ERC practices. Res. Eval. 21, 48–60. https://doi.org/10.1093/reseval/rvs001

Marsh, H.W., Jayasinghe, U.W., Bond, N.W., 2008. Improving the Peer-Review Process for Grant Applications: Reliability, Validity, Bias, and Generalizability. Am. Psychol. 63, 160–168. https://doi.org/10.1037/0003-066X.63.3.160

Merton, R.K., 1973. The sociology of science: theoretical and empirical investigations. University of Chicago Press.

National Academy of Sciences, National Academy of Engineering, and Institute of Medicine, 2007. Rising Above the Gathering Storm: Energizing and Employing America for a Brighter Economic Future. Washington, DC: The National Academies Press. https://doi.org/10.17226/11463.

National Academies of Sciences, Engineering, and Medicine, 2017. An Assessment of ARPA-E. Washington, DC: The National Academies Press. https://doi.org/10.17226/24778.

National Institutes of Health, 2017. Budget [WWW Document]. URL https://www.nih.gov/about-nih/what-we-do/budget

National Institutes of Health, 2008. 2007-2008 Peer Review Self-Study.

Rosenberg, N., 1996. Uncertainty and Technological Change, in: The Mosaic of Economic Growth. pp. 334–353.

Rosenberg, N., 1990. Why Do Firms Do Basic Research (With Their Own Money)? Res. Policy 19, 165–174.

Roy, R., 1985. Funding Science: The Real Defects of Peer Review and an Alternative to it. Sci. Technol. Human Values 10, 73–81.

The American Academy of Arts & Sciences, 2014. Restoring the Foundation: The Vital Role of Research in Preserving the American Dream.

Travis, G., Collins, H., 1991. New Light on Old Boys: Cognitive and Institutional Particularism in the Peer Review System. Sci. Technol. Hum. Values 16, 322–341.

Uzzi, B., Mukherjee, S., Stringer, M., Jones, B.F., 2013. Atypical combinations and scientific impact. Science 342, 468–472. https://doi.org/10.1126/science.1240474

Van Noorden, R., 2015. Biochemist questions peer review at UK funding agency. Nature 1–4. https://doi.org/doi: 10.1038/nature.2014.16479

Waldrop, M., 2008. DARPA and the Internet Revolution, DARPA: 50 Years of Bridging the Gap.

Wang, J., Veugelers, R., Stephan, P., 2017. Bias against novelty in science: A cautionary tale for users of bibliometric indicators. Res. Policy 46, 1416–1436. https://doi.org/10.1016/j.respol.2017.06.006

Wessely, S., 1998. Peer review of grant applications: what do we know? Lancet 352, 301–305. https://doi.org/10.1016/S0140-6736(97)11129-1