Uncertainty and Individual Discretion in Allocating Research Funds

Anna P. Goldstein,[1] Michael Kearney[2]

[1] Harvard University, John F. Kennedy School of Government, Belfer Center for Science and International Affairs
[2] Massachusetts Institute of Technology, Sloan School of Management

Working paper version: May 30, 2017

Abstract

The Advanced Research Projects Agency – Energy (ARPA-E) was created in the model of the Defense Advanced Research Projects Agency (DARPA) to pursue high-risk and transformational energy innovation. Both ARPA-E and DARPA emphasize the autonomy of program staff to make decisions, including which proposed projects to fund. To date, there have been no quantitative studies of how the use of individual discretion impacts the research portfolio, compared to more traditional ranking of peer review scores. Using internal program data from 2009 to 2015, we examine the determinants of project selection and the short-term productivity of ARPA-E projects in terms of publishing, patenting, and market engagement. Our key findings are 1) ARPA-E program directors use significant discretion, 2) their choices result in a portfolio with greater collective uncertainty than other selection methods, and 3) there is no evidence that this use of discretion has reduced the quality of ARPA-E's research portfolio.

Keywords: Research funding, project selection, peer review, uncertainty, innovation

1. Introduction

In October 1957, the Soviet Union launched Sputnik 1 into low-Earth orbit, igniting both a space race and an innovation race with the United States. The United States countered with the Apollo Program and a focus on broader innovation efforts, particularly within the Department of Defense (DOD). In 1958, the DOD launched the Advanced Research Projects Agency (ARPA) to make sure that the United States did not suffer any further technological surprises. Later renamed the Defense Advanced Research Projects Agency (DARPA), the agency has grown to maintain an annual budget of $3 billion and is credited with numerous high-profile inventions, including the internet and the Global Positioning System (GPS) (Alexandrow, 2008; Fuchs, 2010; Waldrop, 2008). Though it is seen as an epicenter of significant technological success, DARPA's operations and strategic decisions remain partly shrouded in secrecy. Research on the topic has been limited to narratives and case studies that articulate DARPA's use of active program management and their impact on technology development. Quantitative understanding of how active program management impacts an agency's research portfolio is currently lacking.

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 (The National Academies, 2007). The Advanced Research Projects Agency - Energy (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.

Literature on the DARPA model typically emphasizes a few defining characteristics, including flexible contracting authority, simple hiring processes, flat organizational structure, energetic culture, and program director independence (Bonvillian, 2009; Bonvillian and van Atta, 2011). In our conversations with staff at ARPA-E, it is clear that the latter focus on autonomy to the program director is key, as it underlies the three discrete responsibilities of the ARPA-E program director (PD). First, program directors are tasked with engaging the research community and defining which technical challenges the agency will pursue. Second, program directors are allotted significant discretion in selecting projects for funding. Third, program directors are actively engaged on a quarterly basis with their portfolio awardees, maintaining a constant threat of termination if goals are not achieved. It is not difficult to imagine that each of these individual authorities can have dramatic effects on the portfolio of projects funded by an actively managed program.

Interestingly, the literature on innovation management does not predict that programs seeking to achieve transformational innovation[1] would be structured in this way. Instead, research on innovation strategy suggests that authority should lie with individual scientists (rather than the grant-maker) if the research is intended to be *exploratory*, i.e. based on untested actions with high degrees of uncertainty. Because of the difficulty of contracting for uncertain outcomes (Aghion and Tirole, 1994), exploratory research is best motivated by tolerance for early failure and reward for long-term success (Holmstrom, 1989; Manso, 2011). An often-cited model of investigator independence is the Howard Hughes Medical Institute (HHMI), which funds "people, not projects" and gives its awardees authority to choose their own research direction. Azoulay et al. (2011) found that HHMI awardees were more likely to yield high-impact publications, compared to similarly qualified scientists with project-based funding from the National Institutes of Health (NIH).

In this context, the DARPA model of giving authority to program staff would be expected to incentivize exploitation of known methods rather than exploration of uncertain ideas. And yet, DARPA itself is frequently heralded for its successes in funding high-risk, high-reward research. The idea that the DARPA model is appropriate for stimulating transformational innovation is widespread, as evidenced by the creation of ARPA-E. Moreover, DARPA's practices have in fact shifted over time toward more active management, with concurrent pushback from the academic community over the loss of investigator freedom ("Joint Statement of the Computing Research Community," 2005; Lazowska and Patterson, 2005).

Given this apparent contradiction between scholarly literature on transformational innovation and the popularity of the DARPA model, there is a need for empirical research to measure the effectiveness of the DARPA model's component practices. The need for understanding is especially pressing in the case of ARPA-E, which is only one of several new public institutions attempting to stimulate much-needed energy innovation (Anadón, 2012). Here, we begin this research effort with a descriptive analysis of how one element of active program management—individual discretion in project selection—affects the portfolio of projects selected and the short-term outcomes of those projects. Ongoing work considers separately the impact of program directors' active engagement and interventions in project execution.

Our findings in this paper are threefold. First, ARPA-E program directors do use significant discretion in selecting projects to fund. Second, they use their discretion to fund projects with champions among the external reviewers, without being swayed by skeptics, resulting in a more uncertain portfolio than would

[1] Although the concept of "transformational innovation" is difficult to define, it is fair to say that the creation of ARPANET at DARPA has indeed led to societal transformation.

have been selected by score ranking. Third, we find no evidence that this use of discretion has reduced the short-term productivity of ARPA-E's project portfolio, although it could increase the potential upside for long-term innovation outcomes.

2. Background

2.1. Research Funding Decisions

For decades, the conventional approach to allocating funds from a public research funding program has been to use a peer review process of some kind. For nearly as long, there have been debates within the scientific community about whether and how to use peer review to determine which research ideas are worth funding. 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 ways of organizing peer review and alternative methods of allocating funding, in order to make a clear comparison 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). Since then, a variety of different peer review systems have proliferated. Because the term "peer review" does not refer to a single process, it is difficult to discuss specific elements of a proposal review system using such a broad term. Here, we will use "peer review" to mean any system in which proposed research is evaluated by at least one expert in the subject of the proposal. The definition of "expert" may be debated (indeed, the level of expertise of a given set of peer reviewers has been one area of criticism). Here, we take "expert" to mean an active member of the same research community as the researcher who authored the proposal; expert reviewers may be internal to the funding program or external.

Expert opinions can be solicited on any number of attributes for a set of proposals. In this paper, we are interested in the decision-making process for measuring each proposal's quality against the chosen criteria. "Quality" is of course a broad term, standing in for several possible review criteria, including originality, scientific merit, feasibility, and usefulness. We use the word to indicate the value or promise that a proposed research project holds for the agency, whatever that agency's goals may be. The criteria used by several US funding programs, specifically along the dimensions of applicability and scientific contribution, have been reviewed by Gans and Murray (2012).

The most common implementation of peer review, 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 on the basis of their expected outcomes. This group may be assembled in-person or consulted individually in writing. An in-person panel 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 any 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 variations of peer review in place 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).

Complaints about peer review may be broadly categorized as pertaining to either the efficiency or the effectiveness of the system (Guthrie et al., 2013; Ismail et al., 2009). Foremost among the efficiency concerns is that, when funding is highly constrained, scientists may be wasting valuable research time applying for grants and reviewing applications with a low success rate. Another category of complaints is about the effectiveness—the ratings of peer reviewers may not provide reliable information on the quality of a proposal, particularly when the idea is novel (American Academy of Arts & Sciences, 2008; Boudreau et al., 2016; Luukkonen, 2012). 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)

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).

Much of the quantitative work that has been done on peer review focuses on US biomedical research support, specifically in the NIH (perhaps due to its size, longevity or willingness to make data accessible to researchers). Li and Agha found informational 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 its number of

publications and citations (Li and Agha, 2015). Yet follow up work by Fang and co-authors on the same dataset found that scores were only able to predict performance at the top percentiles and not for the majority of grants (Fang et al., 2016). Lauer and co-workers studying funding at the National Heart, Lung, and Blood Institute (one of the institutes at NIH) found no associations between percentile score and R01 grant productivity when accounting for grant amount (Lauer et al., 2015).

Reliance on an individual expert opinion to select proposals, as opposed to a more "democratic" system of averaging many 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 directors may seek external opinions but are not bound to act on them. Individual discretion is promoted 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). Others, however, have called for a greater number of reviewers to provide greater "statistical precision" in determining a proposal's value (Kaplan et al., 2008). And in practice, a funding program that does not select projects with the highest mean scores from a group of external reviewers 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 staff to choose which projects to fund. This practice allows us to compare the decisions made by these staff members to the alternative decisions that could have been made based on the external peer reviews.

2.2. ARPA-E

ARPA-E was established by the America COMPETES Act in 2007 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). The first solicitation from ARPA-E stated its intention to fund "high-risk concepts with potentially high-payoff" (ARPA-E, 2009).[2]

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. ARPA-E states that their goal is to support research that "creates fundamentally new learning curves" (ARPA-E, 2015). There

[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."

is no clear path or roadmap to successfully creating or enabling new technology. Uncertainty is therefore a desirable feature for ARPA-E's operations, due to the uncertain nature of transformational innovation compared to research that pursues incremental advances to existing technology.

ARPA-E's funding cycles begin with the hiring of a program director (PD) for a three-year term. At the start of their tenure, program directors are tasked with designing their own technical program. They are given an initial period after hiring to explore ideas for new "white space." They are hired for their technical expertise, but they are not confined to any particular research topic. They are expected to engage with their target research communities and host stakeholder gatherings to refine their ideas.[3] Program directors then pitch their program to ARPA-E leadership, who either accept the idea or encourage further exploration. This process can take up to 18 months of a 36 month contract, and concludes with the release of a funding opportunity announcement (FOA) authored by the program director.

One example of a FOA is Batteries for Electrical Energy Storage in Transportation (BEEST) issued in 2010, which 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.

After the release of a FOA, the program director then oversees the merit review process.[4] The first stage of proposal review at ARPA-E is the submission of concept papers, which contain 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. Applicants are then provided with reviewer comments and are given the chance to briefly

[3] ARPA-E periodically issues "open" solicitations that are open to all areas of energy technology, as opposed to the "focused" programs in a particular technology area that we describe here.

[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. The following summary of the proposal and selection process is based on an example Merit Review Plan provided by ARPA-E, as well as discussions with ARPA-E staff.

reply to these comments. At the end of the review process, the PD submits a recommendation to the Director of ARPA-E of which proposals to select.[5] The recommendations are based on a PD's own review of the application, the content of the external reviews, and the replies received from the applicant. Some proposals are then selected by the Director for negotiation to become a funded project. Our discussions with ARPA-E staff suggest that a majority of selection decisions follow the recommendation of the PD.

It is clear from the above description of ARPA-E's selection procedures that PDs are nominally empowered to use their judgment and discretion to allocate funding. Yet the reality of how PDs use this empowerment could vary dramatically. Perhaps ARPA-E funds only those projects that are well-liked by reviewers, despite giving PDs the ability to make choices independent of review scores. Or perhaps ARPA-E instead funds highly controversial projects, consistent with their stated goal of pursuing high-risk ideas. By examining their proposal review scores and selection data, we will establish how individual discretion is implemented in practice at ARPA-E. We will then examine project outcomes in order to determine whether projects selected using discretion perform differently in the short term.

3. Data

Over the course of two on-site visits to ARPA-E, we compiled datasets on the review of all full proposals and the management of all projects in ARPA-E's funding history. 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 data were completely scrubbed of identifying information in order to protect the confidentiality of the applicants.

3.1. Proposals

Our dataset of proposals contains all review scores for 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:[6]

1. Impact on ARPA-E Mission Area
2. Overall Scientific and Technical Merit

[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] Review questions for OPEN 2012, 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.

3. Qualifications, Experience and Capabilities
4. Sound Management Plan[7]

We used the weights stated in the FOA for each component (Impact, Merit, Qualifications, and Management) to calculate an overall score for each proposal-reviewer pair.[8] 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; we analyze the numerical review scores, which are not revised to reflect this new information.

For the purpose of understanding decision-making by an individual PD, we exclude projects from "open" (non-targeted) programs, for which decision-making around projects 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 only scores from only one external reviewer, so these are excluded from any analysis of standard deviation around a 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 for negotiation.

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

Variable	N	Mean	S.D.	Min.	Max.
Selected	1216	0.3	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 listed in Section 3.1.

3.2. Projects

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

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

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.

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.

We create an indicator variable for whether a project was selected based on individual discretion, compared to a counterfactual set of ranking methods. We call these projects "promoted," in the sense that they were selected despite a relatively low score. Our general method for identifying "promoted" projects is to create a hypothetical score cutoff for each program; this is the cutoff that would be used if projects were selected for funding based on ranking review scores.

We create the 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.[9]

In order to address the impact of project selection practices, 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. Furthermore, 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 ARPA-E projects.

[9] 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 all award or work authorization numbers for ARPA-E projects. 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.[10] 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.

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.

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

Table 2: 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

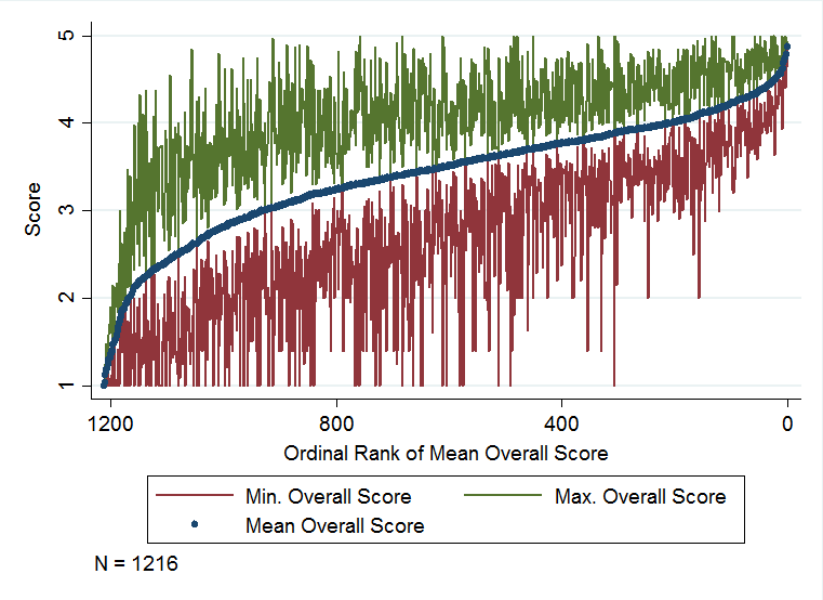
Note: Sample is the set of ARPA-E projects completed 2009-2015 (N = 165) within targeted research programs with overall weighted review scores in the format listed in Section 3.1. The "promoted" variable marks whether a proposal was selected despite a score below a hypothetical cutoff, calculated based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

4. Results

4.1. Distribution of Review Scores

In this section, we describe trends in the selection decisions at ARPA-E with respect to external review scores. One of the scoring elements we investigate is the extent of disagreement among reviewers, as measured by the standard deviation of scores; standard deviation has been used elsewhere as a measure of risk or volatility for a proposed project (Linton, 2016). We also consider the extremes (minimum and maximum) of the score distribution, to account for asymmetric disagreement on either side of the mean score. Reviewers frequently disagreed on the score of an ARPA-E proposal, as shown in Figure 1. The standard deviation of scores around the mean for a given proposal ranges from 0 to 2.6.

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 point of comparison for ARPA-E's selection decisions, we will consider the counterfactual scenario where external peer review scores are the sole determinant of selection. In fact, there are multiple counterfactuals, corresponding to the multiple ways that a group of scores could be ranked. We consider three potential methods of sorting proposals in order of quality, based on a group of scores:

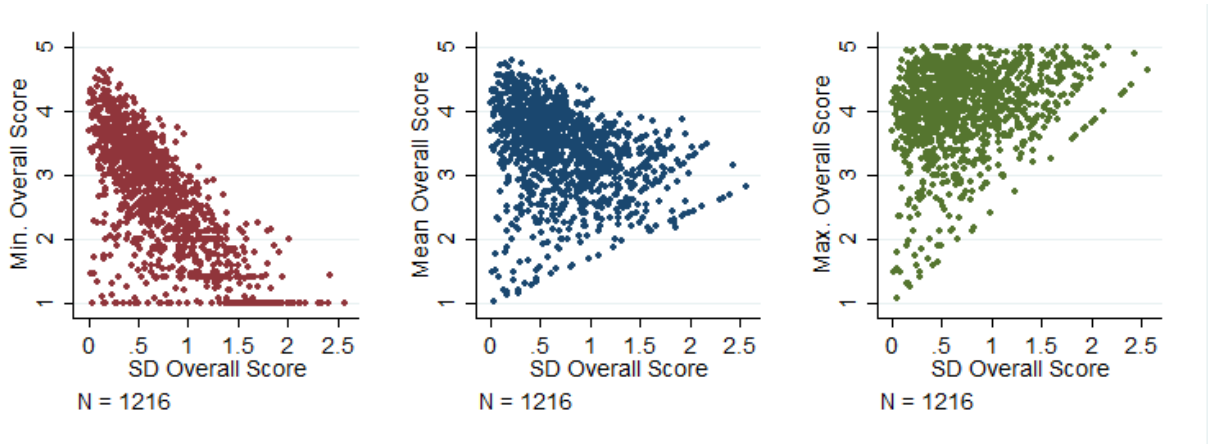
1. *Rank by mean score*: weighs every score equally in determining a proposal's quality
2. *Rank by minimum score*: selects for proposals with no detractors, by only weighing the lowest score received by each
3. *Rank by maximum score*: selects for proposals with champions, by only weighing the highest score received by each

Score-ranking methods are most relevant to the ARPA-E context, and yet these three methods are also analogous to three methods of decision-making in a panel setting: (1) majority vote, (2) consensus, and (3) championing.

Importantly, each of the three score ranking methods has a different implication for the extent of uncertainty in the funded portfolio. This concept is illustrated in Figure 2, where the extent of possible disagreement decreases as the mean score approaches its limit of 5.0. Selecting proposals with the highest mean scores will therefore place an upper limit on the uncertainty of the resulting set of projects. The effect is even stronger for minimum scores; selecting the highest minimum scores mechanically limits 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 be scored with a maximum of 5.0, even if all but one reviewer scored it a 1.0.

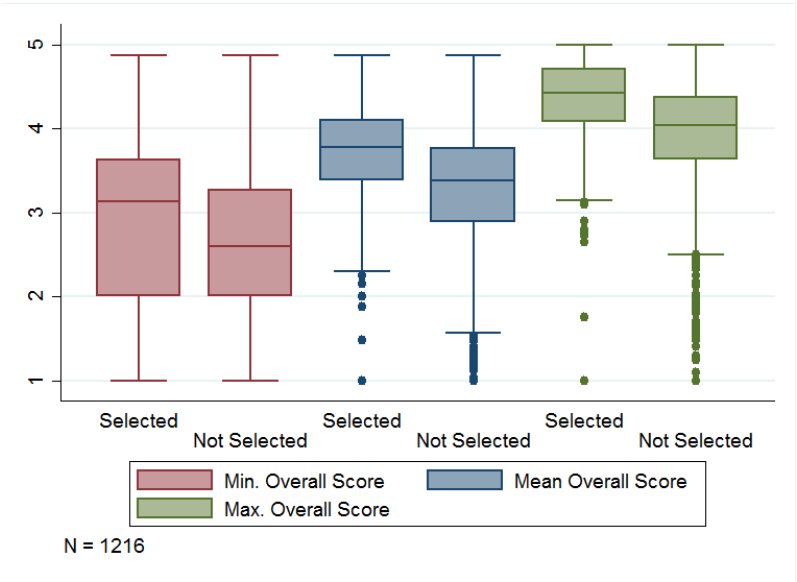
Figure 2: Scoring Statistics for ARPA-E Proposals



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? While PDs are apparently not required by ARPA-E policy to fund the highest scoring proposals, they may still have chosen to do so. Comparing the overall scores for proposals that were funded to those that were not funded, it is clear that neither mean, nor minimum or maximum score 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

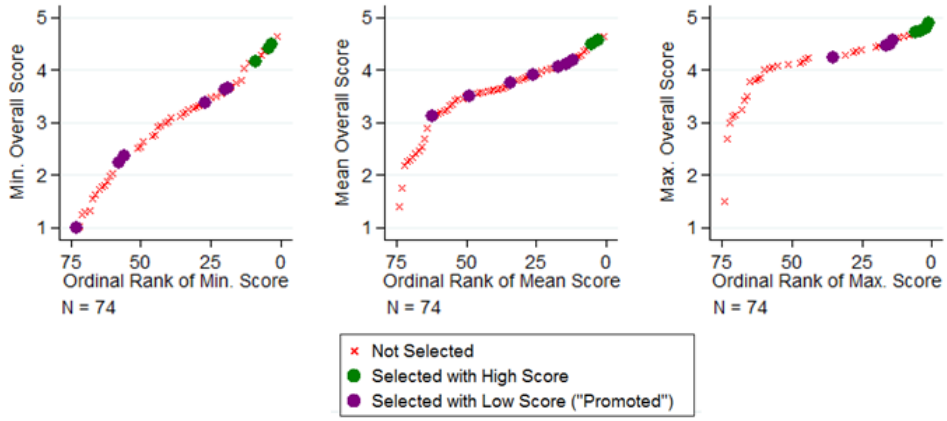


Note: Modified box plot depicts percentiles of score distributions for funded and unfunded ARPA-E proposals. Outside values (< 25th percentile – 1.5 · interquartile range) are plotted as points.

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. In order to analyze the extent of program director discretion used to select ARPA-E projects, we consider selection at the level of the technical program.

Comparing the scores of funded and unfunded proposals in a single technical program (Figure 4), we see that ARPA-E clearly did not use any of the three systematic selection methods by itself. The fact that the set of selected projects cannot be inferred based solely on their scores is the hallmark of individual discretion. This result is confirmed by the proportion of selected projects across the entire dataset that are labeled "promoted" by each counterfactual selection method: 55% "promoted" by mean score, 52% "promoted" by minimum score, and 47% "promoted" by maximum score (Table 2).

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.

Despite the clear evidence of individual discretion for project selection at ARPA-E, it is likely that selection decisions correlate with the external review scores in some way. After all, ARPA-E program directors are technical experts and members of the same research community from which external reviewers are sourced. If scores were unrelated to selection, then we would expect to observe a higher percentage of projects labeled as "promoted"—closer to the overall rejection rate for full proposals, which is 69%.

4.2. Determinants of Selection

Having shown that ARPA-E program directors use individual discretion for project selection, rather than algorithmic score ranking, we now ask: which element(s) of the score distribution (e.g. center of mass, width, one of the two tails) correlate with ARPA-E selections? Are ARPA-E selection decisions at all predictable based on these characteristics? We estimate the correlations using the linear probability model in Equation 1.

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

Y_i is the binary outcome variable for whether proposal i was selected; $Score_i$ is the variable 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. Results using a logit model are shown in the Appendix.

Of the three measures available for ranking (min., mean, and max.), mean score is the most predictive of selection; there is a 19% greater probability of selection for each additional point in the mean overall score (Table 3). Yet the R^2 value is low (0.13), indicating that only a small portion of variation in selection is explained by the mean score. Both minimum score and maximum score are predictive as well, although 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 3: 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. We also qualify the strength of the correlations between score and selection, by comparing to the relationship between score and hypothetical selection by each of the three systematic selection methods. We find that the linear coefficients predicting selection for both mean score and maximum score (0.194 and 0.174, respectively) are less than half what they would be if ARPA-E selected projects by ranking those scores (0.468 and 0.415). 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.

Knowing that individual discretion has the potential to increase ARPA-E's exposure to uncertainty, we estimate the predictive power of reviewer uncertainty 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, but it has a positive and significant coefficient when controlling for the mean overall score. In other words, ARPA-E PDs tend to fund proposals on which reviewers disagree, given the same mean overall score. When minimum and maximum score are included in the same model, the coefficient on minimum score disappears. This suggests that ARPA-E PDs are more likely to select 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 3). 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 in Table 3. The findings above are robust, except that the coefficient on standard deviation in Model 5 loses significance. Yet the coefficient on maximum score in Model 7 gains in both size and significance. Rather than simply selecting projects that have a wide spread of scores, ARPA-E PDs tend to select projects specifically with *upside potential*, indicated by the presence of a high rating by at least one reviewer.

Table 4: Spread of Review Scores for "Promoted" and "Demoted" Proposals

Dependent Variable: Standard Deviation of Overall Review Scores	(1)	(2)	(3)
"Promoted" (low mean score)	0.198**		
	(0.076)		
"Demoted" (high mean score)	-0.224***		
	(0.038)		
"Promoted" (low min. score)		0.292***	
		(0.074)	
"Demoted" (high min. score)		-0.327***	
		(0.033)	
"Promoted" (low max. score)			-0.003
			(0.065)
"Demoted" (high max. score)			0.198***
			(0.035)
Program F.E.	Y	Y	Y
N	1173	1173	1173
R^2	0.238	0.315	0.202

Notes: Standard errors in parentheses. All regressions are OLS with robust standard error, clustered by technical program. "Promoted" proposals were selected for funding from below a hypothetical score cutoff; "demoted" proposals were rejected from above the same cutoff.
* p < 0.10, ** p < 0.05, *** p < 0.01

For an additional perspective on uncertainty, we compare the spread of scores across three groups of proposals: those that were "promoted" from low scores, those that were selected for funding with high scores, and those that were rejected despite high scores—we call this third category of proposals "demoted", as a corollary to "promoted". Compared to a counterfactual selection method of funding only proposals with the highest scores, the effect of individual discretion is to replace "demoted" proposals with "promoted" ones.

The results show that there is greater *ex ante* uncertainty in the ARPA-E research portfolio compared to proposals with the highest mean scores (Model 1). "Promoted" proposals with a low mean score have a wider spread of scores (standard deviation of 0.97 on average), and "demoted" proposals with a high mean score have a narrower spread (standard deviation of 0.55 on average). The same trend appears for ranking by minimum scores (Model 2). Interestingly, selecting projects based on maximum scores would have actually increased the uncertainty in ARPA-E's portfolio (Model 3). This echoes the finding in Figure 2—championing allows by far the greatest extent of uncertainty in project selection of the three score ranking methods.

4.3. Short-Term Impact of Selection Method

In the previous section, we arrived at an empirical understanding of how ARPA-E program directors make selection decisions. In this section, we attempt to establish the effect of these choices on the performance of ARPA-E's research portfolio. Using the output metrics available for the set of completed projects 6 years after the formation of ARPA-E, we make an early assessment of the productivity of ARPA-E's portfolio.

The outputs measured here are directly tied to the ARPA-E award number. Because we do not observe outputs from unfunded applicants, in this paper, we do not attempt to identify the causal effect of ARPA-E funding on the total volume of outputs (e.g. patent applications) from the portfolio as a whole. Instead, we measure the impact of ARPA-E's selection method, recognizing that the "promoted" projects would likely not have received funding from other agencies that rank peer review scores.

Modeling the probability of external 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. External outputs may be associated with both the organization type (university, for-profit, etc.) 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 use of individual discretion by ARPA-E PDs results in greater or lesser productivity, in terms of publications, patent applications, or market engagement metrics, compared to the three alternative selection methods based on score ranking. In other words, is there an association between the "promoted" variable and any of the external metrics? We address this question with regressions of the form shown in Equation 2:

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

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 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 5, we test several models for the relationship between "promoted" (on the basis of mean scores) and whether or not a project produced a publication. No association is found, regardless of the control variable structure.

Table 5: 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 6 show that we do not measure a 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 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 6, there were 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 run the same 15 regressions using (i) mean, min. and max. review scores and (ii) an alternative calculation of the "promoted" variable based on the program budget. Again, we find no consistent trends.

Table 6: 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.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

As a further check for differences in project performance among proposals with different scores, we create two additional variables: one for being "promoted" by 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 all three ranking methods. Including these two variables in the same regressions (Table 7) further demonstrates that the projects selected via PD discretion have roughly equivalent performance to those that were uncontroversial (i.e. scored very highly in external review).

Table 7: Outputs of "Promoted" All Around vs. Not at All "Promoted"

	(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.
* p < 0.10, ** p < 0.05, *** p < 0.01

Beyond the five metrics above, we also consider other ways to measure a project's impact: 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 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 (1) that ARPA-E program directors employs individual decision-making in project selection, rather than any particular ranking method, (2) that this discretion leads them to accept projects with greater uncertainty reflected in reviewers' opinions, and (3) that projects "promoted" from a low score perform comparably to the high-scoring projects in terms of short-term outputs.

Regarding the first key result, nearly 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. Selection decisions do correlate with mean scores, specifically on the proposal's Impact and Merit, but this correlation is much lower than it would be for a program that selected projects based on a ranking of external review scores. ARPA-E PDs choose external reviewers and assign them to specific proposals, at which point they can interpret scores alongside comments, as well as their knowledge of that reviewer's experience and expertise. This process could in fact allow ARPA-E selections to reflect reviewer opinion more accurately than the use of ranked average scores, given the possibility that each reviewer could have different habits and preferences with regard to ratings scales.

We note an important caveat: our findings describe correlations, rather than claiming a causal effect of scores on selection. Program directors may be influenced causally by external reviews to make certain decisions, or they may base their decisions entirely on unobserved variables that happen to correlate with external reviews. In either case, our measurement of correlations is valid for the purpose of describing the features of selected projects, and therefore ARPA-E's research portfolio as a whole. We do not attempt to describe the mindset of each program director when making selections, as this certainly varies widely between individuals.

In the second key result, proposals with high maximum scores are more likely to be selected, controlling for median score. This indicates that the research portfolio constructed by individual program directors carries more collective uncertainty than one based on traditional peer review by score ranking. Selected projects tend to have champions, i.e. someone who rated the proposal very highly. The champion-based selection method is used elsewhere, such as the Gates Foundation's Grand Challenges Exploration programs (Grand Challenges, 2016) and some angel financing groups (Kerr et al., 2014). Importantly, minimum score was found to have no significant relationship with selection at ARPA-E, when accounting for median score. Based on the trends depicted in Figure 2, ranking by minimum score would place the most severe restriction on reviewer disagreement, resulting in a minimally uncertain portfolio of funded projects. The dissimilarity between this method and ARPA-E's selection practices indicates an openness to uncertainty on the part of ARPA-E PDs.

Returning to the quandary posed in the Introduction, it appears that individual discretion has not discouraged the funding of uncertain, exploratory research at ARPA-E. Rather, the expectation that individual discretion would allow greater uncertainty upfront is bolstered by our findings. We suggest two reasons that individual discretion has been particularly effective at encouraging uncertain research directions in the case of ARPA-E. First, ARPA-E's mission is to pursue technological transformation, which requires highly novel solutions. The agency intentionally aims its efforts toward "white space", areas that are inherently less well-established, and as such, less likely to receive consensus approval from the research community. Second, ARPA-E is able to hire technical experts as program directors for short-term rotations, during which they are empowered to create and manage a funding program in their area of expertise.[11] With the opinions of external reviewers as inputs, ARPA-E program directors are able to balance the agency's emphasis on uncertainty with their own expert judgement.

[11] The original authorizing act for ARPA-E specifies a 3-year renewable term and the authority of the Director to hire personnel "without regard to the civil service laws" (110th Congress, 2007).

It is important to note that are no incentives tied to short-term project performance metrics for the PDs, whose employment with the agency is relatively brief—the standard term is 3 years, which is also the average duration of a single project. A PD who is motivated instead by the long-term possibility of impact is likely more willing to make controversial choices. Indeed, interviews with ARPA-E staff indicate that there is indeed a culture of risk-taking at the agency, which incentivizes program directors to select some proposals that are not uniformly liked by external reviewers. Furthermore, within each technical program, PDs are encouraged to construct a diverse portfolio along the dimensions of risk and technology, in order to maximize the chance that one or more projects will be able to achieve the targets set out in the FOA. An example can be seen in the projects funded by the BEEST program (described above in Section 2.2), which included new anode materials, manufacturing processes, and non-lithium battery designs. The value of a diverse portfolio is not captured by review scores for individual projects and can only be taken into account by a PD with a holistic view of the program.

Our final key result illuminates the impact of individual decision-making on the early outcomes of the ARPA-E portfolio. We measure very few differences between low-scoring and high-scoring projects across any metric of impact, indicating the limited informational value in the numeric scores given by ARPA-E's external peer reviewers. This result is consistent with findings that review scores are poorly predictive of grant productivity at the NIH (Fang et al., 2016), although with less precision due to the much smaller sample of completed ARPA-E awards. The 95% confidence intervals for the 15 coefficients shown in Table 6 range from -0.28 to 0.27; the average upper and lower limits are 0.15 and -0.16 respectively. We can say with relative confidence that there is no more than a 28% reduction in the probability of short-term outputs for "promoted" projects, and indeed there could be up to 27% increase in probability.

At this early stage of assessment, it seems as though ARPA-E used individual discretion to compile a portfolio of projects with greater uncertainty, which helps fulfill their mission to pursue "white space," without seeing significant downsides in terms of the short-term productivity of those projects. Because our study design does not capture outcomes 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 as a whole 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 both patenting and publication outcomes compared to other funding sources within DOE (The National Academies 2017).

Publications, patent applications and market engagement are not measures of success in themselves, but they mark 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 (and to measure) as the technologies created with ARPA-E support are developed and deployed. It is plausible that projects selected using individual decision-making at ARPA-E have more divergent long-term outcomes, due to a higher level of uncertainty *ex ante*. This idea is illustrated by the ARPA-E mantra, "If it works, will it matter?" Even if many ARPA-E projects result in technical failure, the costs of those projects could be overwhelmed by the returns on just a few hugely impactful projects, resulting in overall greater long-term productivity of the agency's investments. 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 (GPS) (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.

6. Conclusion

One facet of the DARPA model—empowering program staff to use discretion in allocating research funds—is quantitatively described here for the first time. Using data on project selection at ARPA-E, we have shown how program directors use significant autonomy to select proposals that would not have been selected based on external review scores. It is worth re-iterating that the fact of program directors' empowerment at ARPA-E does not itself imply this outcome for the agency. Empowered PDs could have chosen instead to fund projects that overlap exactly with the opinions of external reviewers, thus limiting the amount of uncertainty the agency takes on board. Our findings here point to an intentional choice on the part of the agency, in hiring PDs or training them or both, to encourage selection of uncertain projects for the sake of pursuing transformational energy research.

We find that PDs prefer to fund proposals with champions, given a certain mean review score. This tendency results in greater *ex ante* uncertainty in the ARPA-E portfolio, and we observe no significant detriment to the agency's short-term research productivity as a result. In other words, as implemented at ARPA-E, individual discretion appears no better or worse than algorithmic score ranking. And yet there may be an advantage in the long-term productivity of the agency, if proposals with more divergent reviewer opinions experience more divergent outcomes, i.e. more failures *and* more successes with transformational impact. This thread of research will benefit from the passage of time, with a larger sample of projects and the ability to measure longer-term outcomes.

The findings in this paper rely on a counterfactual scenario where some of the proposals that were selected would have been otherwise rejected; unfortunately, we are not able to make any claims about

those proposals that were rejected and would have been otherwise selected. This is a typical situation for any program whose stakeholders are resistant to the idea of running randomized controlled trials of different project selection practices. By collecting external peer reviews alongside their implementation of individual discretion, ARPA-E has provided a unique window into the relative effectiveness of their approach to project selection.

Acknowledgements

Our analysis originated as a consulting engagement with the National Academies of Science, Engineering and Medicine for a study on ARPA-E (The National Academies 2017). 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, Josh Krieger, Gilbert Metcalf, Ramana Nanda, Venky Narayanamurti, Scott Stern, and participants in the NBER Productivity and Innovation seminar. We also thank current and former ARPA-E staff, in particular Dave Dixon, Ron Faibish, Andy Kim and Ashley Leasure, for their assistance in data collection. All errors or omissions are our own.

References

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

Aghion, P., Tirole, J., 1994. The Management of Innovation. Q. J. Econ. 109, 1185–1209.

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

American Academy of Arts & Sciences, 2008. Advancing Research In Science and Engineering: Investing in Early-Career Scientists and High-Risk, High-Reward Research.

Anadón, L.D., 2012. Missions-oriented RD&D institutions in energy between 2000 and 2010: A comparative analysis of China, the United Kingdom, and the United States. Res. Policy 41, 1742–1756. doi:10.1016/j.respol.2012.02.015

ARPA-E, 2010. BEEST Program Overview.

Azoulay, P., 2012. Research efficiency: Turn the scientific method on ourselves. Nature 484, 31–32. doi: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.

Bollen, J., Crandall, D., Junk, D., 2014. From funding agencies to scientific agency. EMBO Rep. 15, 1–3.

doi:10.1002/embr.201338068

Bonvillian, W.B., 2009. The Connected Science Model for Innovation—The DARPA Role, in: 21st Century Innovation Systems for Japan and the United States: Lessons from a Decade of Change. National Academies Press.

Bonvillian, W.B., van Atta, R., 2011. ARPA-E and DARPA: Applying the DARPA model to energy innovation. J. Technol. Transf. 36, 469–513. doi: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. doi:10.1287/mnsc.2015.2285

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

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. doi:10.1287/mnsc.1040.0290

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

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. doi:10.1016/j.respol.2010.07.003

Gans, J.S., Murray, F.E., 2012. Funding Scientific Knowledge: Selection, Disclosure, and the Public-Private Portfolio, in: Lerner, J., Stern, S. (Eds.), The Rate and Direction of Inventive Activity Revisited. University of Chicago Press, pp. 51–103.

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

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

Holmstrom, B., 1989. Agency costs and innovation. J. Econ. Behav. Organ. 12, 305–327. doi:10.1016/0167-2681(89)90025-5

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. doi:10.1073/pnas.0804538105

Joint Statement of the Computing Research Community, 2005. . House Sci. Comm. Hear. Futur. Comput. Sci. Res. U.S.

Kaplan, D., Lacetera, N., Kaplan, C., 2008. Sample size and precision in NIH peer review. PLoS One 3, 3–5. doi: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. doi:10.1093/rfs/hhr098

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

Lazowska, E.D., Patterson, D.A., 2005. An endless frontier postponed. Science 308, 757. doi:10.1126/science.1113963

Li, D., Agha, L., 2015. Big names or big ideas: Do peer-review panels select the best science proposals? Science 348, 434–438. doi: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. doi: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. doi:10.1093/reseval/rvs001

Manso, G., 2011. Motivating Innovation. J. Finance 66, 1823–1860. doi:10.1111/j.1540-6261.2011.01688.x

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. doi:10.1037/0003-066X.63.3.160

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

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.

The National Academies, 2007. Rising Above the Gathering Storm: Energizing and Employing America for a Brighter Economic Future.

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

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

Appendix

Table A8: Predicting Selection by Review Score Distribution – Logit Model

Dependent Variable: ARPA-E Selected Proposal for Funding	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Mean Overall Score	3.415***				6.480***		
	(1.130)				(1.743)		
Min. Overall Score		1.507*				1.155	0.853
		(0.351)				(0.240)	(0.170)
Max. Overall Score			3.410***			3.196***	1.750**
			(0.880)			(0.697)	(0.414)
SD Overall Score				1.358	3.077***		
				(0.474)	(1.263)		
Med. Overall Score							2.620***
							(0.532)
Program F.E.	Y	Y	Y	Y	Y	Y	Y
N	1216	1216	1216	1173	1173	1216	1216
Pseudo R^2	0.120	0.067	0.117	0.052	0.163	0.118	0.137

Notes: Standard errors in parentheses. All regressions are logit with robust standard error, clustered by technical program. Coefficients are exponentiated, i.e. odds ratio of outcome for two groups with a difference of 1 score unit in the independent variable.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table A9: Predicting Selection by Review Score Distribution – Proposals with >2 External Reviews

Dependent Variable: ARPA-E Selected Proposal for Funding	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Mean Overall Score	0.251***				0.263***		
	(0.039)				(0.036)		
Min. Overall Score		0.101**				0.044	-0.011
		(0.048)				(0.049)	(0.050)
Max. Overall Score			0.216***			0.195***	0.091**
			(0.030)			(0.032)	(0.039)
SD Overall Score				0.034	0.105		
				(0.084)	(0.074)		
Med. Overall Score							0.161***
							(0.034)
Program F.E.	Y	Y	Y	Y	Y	Y	Y
N	943	943	943	943	943	943	943
R^2	0.168	0.106	0.150	0.082	0.174	0.154	0.175

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

Table A10: Predicting Selection by Component Scores

Dependent Variable: ARPA-E Selected Proposal for Funding	(1)	(2)	(3)
Min. Impact Score	0.054*		
	(0.030)		
Min. Merit Score	0.075**		
	(0.034)		
Min. Qualifications Score	-0.049		
	(0.030)		
Min. Management Score	0.012		
	(0.012)		
Mean Impact Score		0.103***	
		(0.030)	
Mean Merit Score		0.147***	
		(0.045)	
Mean Qualifications Score		-0.025	
		(0.040)	
Mean Management Score		-0.021	
		(0.021)	
Max. Impact Score			0.068***
			(0.023)
Max. Merit Score			0.104***
			(0.032)
Max. Qualifications Score			0.026
			(0.022)
Max. Management Score			-0.023
			(0.027)
Program F.E.	Y	Y	Y
N	1216	1216	1216
R^2	0.096	0.156	0.141

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

Table A11: Predicting Counterfactual Selection by Scores

Dependent Variable:	High Mean Score	High Mean Score	High Minimum Score	High Minimum Score	High Maximum Score	High Maximum Score
	(1)	(2)	(3)	(4)	(5)	(6)
Mean Overall Score	0.468***	0.470***				
	(0.033)	(0.040)				
Min. Overall Score			0.388***	0.413***		
			(0.025)	(0.029)		
Max. Overall Score					0.415***	0.468***
					(0.037)	(0.044)
SD Overall Score		-0.142***		0.059		0.004
		(0.029)		(0.038)		(0.026)
Program F.E.	Y	Y	Y	Y	Y	Y
N	1216	1173	1216	1173	1216	1173
R^2	0.461	0.485	0.579	0.577	0.406	0.434

Notes: Standard errors in parentheses. All regressions are OLS with robust standard error, clustered by technical program. The models include a fixed effect for technical program.
* p < 0.10, ** p < 0.05, *** p < 0.01

Table A12: Project Outputs vs. Scoring Element

Dependent Variable:	At Least 1 Publication	At Least 1 Patent Application	Market Engagement	Any External Output	All External Outputs
	(1)	(2)	(3)	(4)	(5)
Min. Overall Score	0.047	0.019	-0.038	0.040	0.011
	(0.041)	(0.027)	(0.029)	(0.026)	(0.031)
N	165	165	165	165	165
R^2	0.371	0.327	0.288	0.366	0.163
Mean Overall Score	0.076	0.075	-0.049	0.059*	0.031
	(0.059)	(0.057)	(0.048)	(0.033)	(0.036)
N	165	165	165	165	165
R^2	0.372	0.333	0.287	0.366	0.166
Max. Overall Score	0.017	0.102**	-0.008	0.022	0.022
	(0.037)	(0.041)	(0.047)	(0.037)	(0.015)
N	165	165	165	165	165
R^2	0.366	0.344	0.284	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 initial award amount, as well as a fixed effect for technical program and a fixed effect for the organization type.
* p < 0.10, ** p < 0.05, *** p < 0.01

Table A13: Outputs of "Promoted" Projects by Budget Criteria

Dependent Variable:	At Least 1 Publication	At Least 1 Patent Application	Market Engagement	Any External Output	All External Outputs
	(1)	(2)	(3)	(4)	(5)
"Promoted" by budget	-0.015	-0.081	0.074	-0.087	0.035
(low min. overall score)	(0.090)	(0.095)	(0.066)	(0.087)	(0.048)
N	165	165	165	165	165
R^2	0.366	0.331	0.288	0.367	0.165
"Promoted" by budget	-0.123	0.002	0.001	-0.025	-0.018
(low mean overall score)	(0.072)	(0.101)	(0.081)	(0.097)	(0.060)
N	165	165	165	165	165
R^2	0.375	0.326	0.284	0.361	0.163
"Promoted" by budget	0.040	-0.123	0.019	-0.035	0.055
(low max. overall score)	(0.062)	(0.077)	(0.078)	(0.081)	(0.047)
N	165	165	165	165	165
R^2	0.367	0.338	0.284	0.362	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.

Table A14: Additional Outputs of "Promoted" All Around vs. Not at All "Promoted" Projects

Dependent Variable:	Number of Publications	At Least 1 Highly Cited Publication	At Least 1 Patent Issued	Private Funding Amount (Million USD)
	(1)	(2)	(3)	(4)
"Promoted" all around (low mean, min., *and* max. overall score)	0.912 (1.131)	0.121* (0.062)	0.063 (0.109)	-4.114 (2.998)
Not "promoted" by any measure	0.564 (0.867)	0.117 (0.086)	-0.043 (0.071)	-1.586 (3.958)
N	165	165	165	165
R^2	0.337	0.208	0.284	0.332

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