Actively Managed Uncertainty and Risk:
Evidence from the First Six Years of ARPA-E

Anna Goldstein, Harvard Kennedy School

Michael Kearney, MIT Sloan School of Management

Draft

September 26, 2016

This analysis originated as a consulting engagement with the National Academies for a study on ARPA-E. We are thankful for helpful discussions with Laura Diaz Anadon, Pierre Azoulay, Paul Beaton, Iain Cockburn, Gail Cohen, Jeff Furman, Gilbert Metcalf, Venky Narayanamurti and Scott Stern. We also thank 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.

Abstract

Research funding programs must attempt to optimize the risk level of their investments, despite the difficulty of assessing risk in the face of the inherent uncertainty associated with knowledge production. We study the use of active program management at the US Department of Energy's Advanced Research Projects Agency – Energy (ARPA-E) and how it has allowed ARPA-E program directors to tailor the agency's exposure to risk and uncertainty. In this context, *active program management* is a set of practices which empowers individuals to select proposals and to manage projects. We empirically demonstrate that ARPA-E has implemented active program management, which has resulted in greater tolerance for uncertainty in project selection, followed by risk mitigation in project management. We also analyze the outputs of ARPA-E projects in publishing, patenting, and market engagement, and we show that their use of active program management is supported by these early indicators of productivity.

1. Introduction

Over the past 60 years, scholars have articulated the importance of public sector support for scientific research to spur knowledge production, accelerate technological change and drive economic growth. Romer (1990) identified technological change as a key enabler of economic growth, and the extant literature highlights two primary reasons for public sector involvement to enhance technological change through research and development. First, private sector firms driven by a profit motive will fail to optimize investment for social return creating a funding gap between private support and the socially optimal level of support for research (Arrow 1962, Griliches 1992, Nelson 1959). Second, the findings of public sector research are more likely to be broadly disclosed, relative to those of private sector research, which is critical for the long-term accumulation of knowledge and economic growth (Romer 1990, Dasgupta and David 1994, Mokyr 2004, Gans, Murray and Stern 2010).

More recently, scholars have turned to the question of how the public sector can best promote technological change. A broad literature explores the impact of intellectual property rights and anti-trust regulations (Scotchmer 1991, Aghion et al. 2001, Aghion et al. 2005, Murray and Stern 2005, Scotchmer 2010, Segal and Whinston 2011, Gans 2011). Others have evaluated the use of grants as subsidies for R&D (Lerner 2000, Howell 2015) and prizes and advance market commitments as alternatives to intellectual property (Kremer and Williams 2010, Murray et al. 2012). Another thread of this literature considers how institutions contract for research (Aghion and Tirole 1994, Maurer and Scotchmer 2004, Gans and Murray 2010, Manso 2011), as well as the productivity of various research management practices (Manso 2011, Azoulay et al. 2011).

Absent from this literature is a serious engagement with the management of risk and uncertainty within public research funding programs. This paper seeks to add to the literatures related to contracting and research management practices with a detailed empirical analysis of the project selection and management strategies at the Advanced Research Projects Agency - Energy (ARPA-E), a grant-making organization within the U.S. Department of Energy.

In the context of research investments, *uncertainty* and *risk* are separate but related concepts. When embarking on an effort to create new discoveries and inventions, a researcher cannot make a complete list of possible outcomes, much less assign probabilities to these outcomes. This stage of innovation is characterized by “true uncertainty,” defined by Knight as *an unmeasurable lack of knowledge* (1921). Risk, on the other hand, represents *a measurable lack of knowledge*, e.g. a coin flip where the outcomes have a known probability distribution, though none of them are 100%.

Risk is an issue in practically every sector of the economy, though uncertainty is especially salient in any discussion of knowledge and technology (Rosenberg 1996). The technological uncertainty inherent to research itself and the uncertainty of application and economic impact combine to make it exceedingly difficult to select the optimal research investments, especially for a mission-driven public agency whose goal is to have a specific impact beyond general economic development.

Rather than giving up on research in the face of uncertainty, firms and government agencies alike have persisted in attempting to identify the research investments that will deliver the greatest returns, whether public or private. The conventional approach to coping with uncertainty in a public research funding program is to solicit expert opinions on the available research proposals, use a consensus process to select the investments, and then allow the researcher to execute the project with minimal involvement by the funders.

An alternative approach, which is our focus here, is for individual program staff members to exercise significant discretion over an agency's research investments. We call this general approach *active program management*. As in an actively managed investment portfolio, an actively managed research program gives individual experts the freedom to select projects from a set of proposals. They are also given the freedom to modify the terms of a project as it proceeds.

Active program management is exemplified by the Advanced Research Projects Agency – Energy (ARPA-E), part of the US Department of Energy (DOE). Since its start in 2009, ARPA-E has pursued high-risk research and development (R&D) for energy technology. ARPA-E was created to be an agile organization, outside the traditional structure and bureaucracy of DOE, and this flexibility has allowed them to adopt active program management, in the style of the Defense Advanced Research Projects Agency (DARPA). Similar to DARPA program managers, ARPA-E's program directors are empowered to use their individual subject matter expertise in the selection and management of a portfolio of research projects.

ARPA-E was evaluated by a committee of the National Academies (National Academies 2016), and we provided expert consultation to the committee. The analysis in this paper is derived from that work, using data obtained from ARPA-E on their history of proposal review and project management.

Our contribution in this paper is twofold. First, we empirically describe the use of active program management strategies in a government research agency. This entails quantitative analysis of how ARPA-E program directors make decisions about project selection and project modification, and the implications of those choices for the agency's approach to risk and uncertainty. Second, we demonstrate the effectiveness of these strategies, in terms of short-term research outputs.

Our key findings are: (1) ARPA-E has internalized tenets of active program management by empowering their program directors to make decisions regarding project selection and management. (2) The use of individual decision-making at ARPA-E has resulted in the selection of projects with greater uncertainty. (3) There is no evidence that accepting greater uncertainty has reduced the productivity of ARPA-E's project portfolio. (4) ARPA-E uses project modifications, such as cutting projects short in response to poor performance, in an attempt to mitigate risk for the agency. (5) There is some evidence that these attempts to mitigate risk have resulted in improved productivity for ARPA-E as a whole.

The paper continues as follows. Section 2 provides background on risk and uncertainty in the context of research investment. Section 3 details ARPA-E's approach to risk. Section 4 outlines the data used in this analysis. Section 5 describes the econometric model used and our findings. Section 6 discusses interpretation of our findings and section 7 offers some concluding statements.

2. Risk and Uncertainty in Research Investment

When time or money is invested in knowledge production, the outcome is unknowable in many dimensions (Rosenberg 1996). We can't know with certainty what information will result from research; we can't know what applications that information will have, and we can't know what the social and economic impact of those applications will be. To be explicit in our vocabulary, we separate *technical uncertainty*, uncertainty of the technical outcome of a research project, from *impact uncertainty*, uncertainty of the project's ultimate social or economic impact. (We also note that these types of uncertainty are not unrelated, as the future impact of a project surely depends on its technical outcome.) For the remainder of this paper, we will use the word uncertainty to refer specifically to technical uncertainty surrounding a proposed research project.

Furthermore, we take care to distinguish *uncertainty* from *risk*, as they relate to a prospective research investment, although they are often used interchangeably. Though the exact definition of "risk" varies depending on the context of the discussion (Holton 2004), common usage indicates that risk is a function of both the impact of a negative event and the likelihood of that event. For the purpose of this paper, the negative event that concerns us is the failure of a project to achieve its technical goals or answer its research questions. In the context of research funding, the impact of a failed investment is mostly limited to the loss of funds. If the variation in investment levels between projects is low enough, then the relative riskiness of an investment is primarily a function of the likelihood that it succeeds. In this work, we use risk to refer specifically to the prospective probability of failure of a research project.

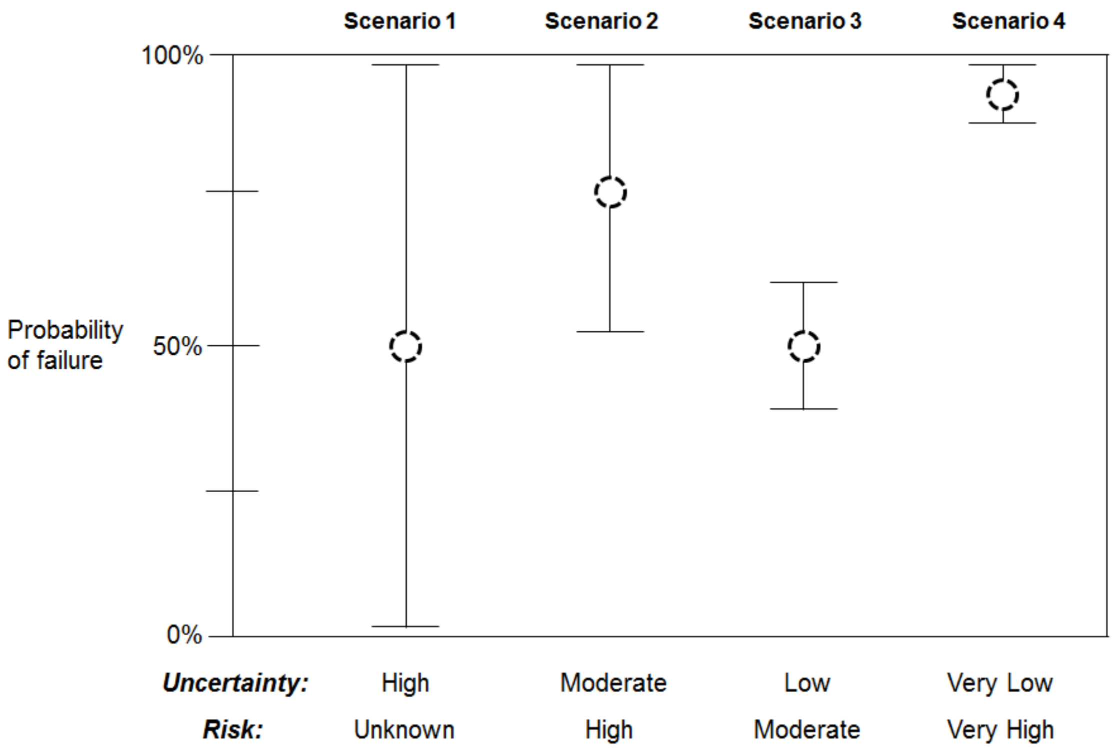
Research investments are commonly discussed in terms of their relative risk. Dietz and Rogers (2012) describe the metaphor for research as a portfolio of financial investments, where the optimal portfolio is one that balances risk. Some government programs specifically aim to fund "high risk, high reward" research. Private companies that invest in research may seek or avoid risk, depending on their business strategy. No matter the desired risk level, any research funding program needs a method for assessing the riskiness of a given research idea.

And so, a paradox arises from the need to assess risk of research proposals, despite the inherent uncertainty associated with R&D investments. When the investment being considered is a research hypothesis—when a project aims to expand the boundaries of human knowledge, either theoretical or practical—there are no historical data available to inform the investor of its prospective risk. Thus, identifying the prospective probability of failure is fundamentally challenging.

Consider an extremely simplified depiction of a research project as an investment that has two possible outcomes: technical success and technical failure. Now consider a poll that asks all the researchers in the specific technical area how likely a particular project is to fail. These experts are expected to have the greatest knowledge and therefore the least uncertainty in a particular technology area. The collective level of uncertainty is determined by the extent of knowledge available, i.e. the sum of knowledge held by the community of researchers in a given technical area.

In a graphical depiction, we can represent uncertainty as the spread of responses to our expert poll for a given project, i.e. the error bars around the prospective probability of failure. If the spread is narrow and the uncertainty is low, then the level of risk can be determined with greater accuracy. Projects with greater degrees of uncertainty are ones in which the prospective risk is harder to ascertain. Figure 1 illustrates this relationship between risk and uncertainty in four hypothetical R&D investment scenarios.

Figure 1: Risk and Uncertainty in R&D Investments



Scenario 1 represents an R&D investment that is characterized by very high uncertainty, indicated by the wide spread of prospective judgments on its level of risk. In this example, the error bars cover the entire probability range for failure. The probability distribution between success and failure is flat, and so the risk level of this investment is truly unknown. In Scenario 2, the error bars are narrower, but still somewhat broad, with the probability of failure between 50 percent and 100 percent. We label this scenario high risk, because the project is judged more likely to fail than to succeed. In Scenario 3, the distribution of judgments narrows even more due to relatively low uncertainty, and we can say that this investment carries moderate risk. Scenario 4 represents a project with high agreement around the high likelihood of failure.

This framework allows for a clear understanding of the difference between risk and uncertainty. A low risk investment must inherently be associated with low uncertainty, because the level of knowledge is at least great enough to judge its probability of failure in the first place. Indeed, these are the investments that would be considered “low-hanging fruit” for a research funding program. However, the relationship between uncertainty and risk does not work in the opposite direction: low uncertainty does not imply low risk. As more knowledge is obtained about an investment, it could instead appear to be prospectively more risky; this is the type of investment depicted by Scenario 4.

When uncertainty regarding a particular technology idea is relatively low, organizations can tailor their investment in a straightforward way to obtain their desired level of risk. However, if a firm or agency decides to pursue an area associated with high uncertainty, how can they make smart decisions about the distribution of research funds? How do they manage the risk associated with their investments, if that risk level is unknown at the outset? These are the fundamental question we address in this paper by examining the strategies of one particular agency, ARPA-E. Before turning to the specifics of ARPA-E, we outline how various program design choices can impact the level of risk and uncertainty in a research portfolio.

2.1. Project Selection to Manage Uncertainty

A typical research funding program solicits proposals, which may be confined to a particular discipline of science or engineering. The program must then select some of these proposals to construct a portfolio of projects, despite the difficulty in judging the underlying risk of ideas with high uncertainty. When faced with this choice, funding programs typically turn to subject matter experts through a peer review system.

These expert members of the research community are least burdened by uncertainty, because they have the greatest extent of information on the research subject. Of course, even experts do not possess perfect information. Each subject matter expert has a slightly different outlook on technology and research ideas based on their experience, and they are likely to disagree with each other about the value of any particular research proposal. A program may consult multiple experts and combine their opinions in some way to arrive at a minimally controversial set of proposals. The phrase “peer review” in research funding typically connotes some sort of consensus, such as the average of numeric ratings received from multiple reviewers.

Consensus peer review has some major benefits for the scientific community. Selecting proposals based on the opinions of expert practitioners of a scientific field limits the opportunity for corruption and political influence in the distribution of public funds. Additionally, using consensus rather than a single reviewer allows agencies to protect against manipulation of funds if that reviewer has a conflict of interest. In its 2014 report *Restoring the Foundation*, the American Academy of Arts & Sciences recommended that “peer review should remain the mechanism by which federal agencies make research award decisions.”

However, alongside the benefits of peer review, there are many known weaknesses in its typical practice, such as the underappreciation of early career scientists’ contributions (Merton 1968) and the discounting of novel ideas by those immersed in their field (Luukonen 2012, Boudreau et al 2016). Meanwhile, there is a debate in the literature over the effectiveness of peer review at selecting the best investments, particularly in a funding environment where the volume of proposals greatly outweighs the available resources. The National Institutes of Health (NIH) uses “paylines” or cutoffs of reviewer scores, below which an application will not be funded (Azoulay et al 2015). Li and Agha (2015) demonstrated that the review score percentile of NIH proposals do correlate somewhat with the productivity of the grant, but this correlation seems to apply only to the top scoring proposals rather than those near the cutoff (Fang et al 2016).

Peer review is expensive for agencies to coordinate and time-consuming for the reviewers. An alternative decision-making strategy, which is implemented under active program management, is to select proposals based on an individual expert’s opinion. The Gates Foundation has used individual decision-making to select research investments in its Grand Challenges Explorations initiative. Individuals may have also been empowered to make funding decisions at NIH, as evidenced by the fact that 13% of funded applications from 2003-2007 were “exceptions” from below the score cutoff (Fang et al 2016).[1]

One possible benefit of empowering individuals to select projects is that it gives a program the opportunity to tailor its level of uncertainty. To explain this, we refer back to the depictions of risk and uncertainty in Figure 1. If the score given by a reviewer correlates with that reviewer’s estimate of the project’s likelihood of success, then the projects with the highest mean scores are those with relatively low uncertainty based on a narrow range of high scores. The wider the disagreement around a proposal, the lower its maximum possible score will be after the

[1] Johnson (2008) explains that “funding decisions [at NIH] are thought to be highly correlated with percentile scores,” though the specifics of the review process varies among different institutes and centers within NIH.

many scores are averaged. If novelty is associated with greater uncertainty, then this is yet another factor that leads consensus peer review to neglect novel ideas.

An individual reviewer's score for a proposal, on the other hand, does not relate inherently to the collective uncertainty of their peers. Making decisions based on an individual's score, therefore, gives the program the option to select proposals with high uncertainty that would have been otherwise eliminated by their lower average scores. In other words, project selection using individual decision-making remains sensitive to the "noise" that would have been smoothed out under consensus peer review. This method directly contradicts those who believe that there should be greater "statistical precision" in peer review (Kaplan 2008).

It should be noted that individual decision-making does not inherently solve the problems associated with consensus peer review. First, an individual reviewer is still subject to psychological effects such as a bias against novel ideas. And second, this strategy does not necessarily lead to the acceptance of novel ideas, because it does not actually *require* the program to adopt greater uncertainty in its research portfolio. The individual in question may in fact choose the same projects that would be chosen under consensus peer review; they may choose an entirely different set of projects, but with an equivalent level of uncertainty.

Individual decision-making in project selection gives a program the option to adopt greater uncertainty. How this strategy actually impacts an agency's approach to uncertainty in practice is a question of implementation. We resolve this question for ARPA-E empirically in Section 5.1.

2.2. Project Modification to Manage Risk

After a round of investments has been selected, the conventional approach to grant-making in the public sector is to allow researchers to manage their own research process independently. Many scholars have described the benefits of investigator freedom. Nelson (1962) recounts the invention of the transistor at Bell Labs as a success story for the practice of giving researchers flexibility to choose research directions. Azoulay et al. (2011) report the impressive outcomes of Howard Hughes Medical Institute (HHMI) awards, which give considerable freedom to the investigator, compared to NIH awards which adhered to pre-defined project aims.

Related to this flexibility and freedom is the practice of tolerating failure. This is another feature of HHMI grants, which are frequently renewed after the first five-year cycle (Azoulay et al. 2011). Tian and Wang (2014) found that tolerance for failure results in more innovation-related productivity within the venture capital industry. And Manso (2011) established that the threat of termination, as opposed to continuation of research funding, encourages researches to exploit well-known actions, as opposed to exploring unknown actions.

One consequence of investigator freedom is that the research program has no opportunity to manage its exposure to risk, i.e. the probability of failure for each of its investments. If a program chooses to continue projects as is without modifications, they accept the risk inherent to its portfolio from the outset. In fact, by forgoing active involvement with the project, grant-makers accept an additional risk as well; there is some chance that as the project continues, the researchers may decide to follow a new direction. This may be acceptable for an agency that aims to advance science in general, though it is likely to be less acceptable for an agency that is pursuing a specific mission, such as energy technology innovation.

An alternative approach to handling ongoing projects, one which we consider part of active program management, gives authority to program staff and empowers them to make decisions throughout the project. Organizations in the private sector frequently use this strategy. Huchzermeier and Loch (2001) details the importance of maintaining flexibility in a firm's approach to R&D risk. In this framework, flexibility entails the use of "real options" for managing a research project, such as delay, abandon, contract, expand and/or switch, all of which are midcourse actions that improve outcomes. Similarly, Trigeorgis (1997) finds that flexibility improves the potential upside of a project while limiting downside losses. Dixit and Pindyck (1994) find that as uncertainty in project payoffs increase, additional measures should be taken by managers to maintain flexibility.

Although the literature about managerial flexibility has focused on the private sector, it could be applied to public sector research management as well. Rather than accept the inherent level of risk associated with a given project after selection, as is largely the case throughout federal government grant-making programs, an agency could choose instead to take a flexible approach and manage the project as it proceeds. ARPA-E does this by enabling its program directors to modify the terms of ongoing projects, as will be discussed in the next section.

Giving an investor the ability to make project modifications allows the project to be adjusted in response to performance. If the investor wants to limit risk, they may choose to deprive failing projects of time and effort. Assuming they receive frequent progress updates from the researchers, then failure to achieve milestones can be identified and acted upon by program staff.

This gives the program as a whole the opportunity to bear lower risk than if all projects were continued at their original level of investment.

Again, as with project selection, empowering individuals to make project modifications does not necessitate that they use their empowerment to mitigate risk. It merely gives the agency the option to affect its exposure to risk, rather than accepting the level of risk present at the outset of each project. In Section 5.2, we will show empirical evidence for how ARPA-E's use of individual decision-making in project modification manifests in terms of risk.

3. Managing Risk and Uncertainty at 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. To this end, ARPA-E was expected to "overcome the long-term and high-risk technological barriers in the development of energy technologies" and to take on projects "in areas that industry by itself is not likely to undertake because of technical and financial uncertainty" (110th Congress 2007, sec. 5012).

The attitude of the agency toward risk was evident in its first funding opportunity announcement (FOA), which states that ARPA-E aims to fund "high-risk concepts with potentially high-payoff" (ARPA-E, 2009). It goes on to say, "The kinds of technologies most suited to an ARPA-E style development are those that still have significant technical risks to overcome, but promise to meet the future costs and scale of products that can deeply penetrate into consumer and industrial use." This makes it clear that ARPA-E aims to fund projects with high technical risk—those that are judged to have relatively low probability of technical success—but high impact if the project succeeds.[2] ARPA-E funding allows researchers to demonstrate their technology idea, so that the ideas with large impact don't have to die on the vine.

Other statements help us understand ARPA-E's approach to risk mitigation. Recent FOAs emphasize "the extent to which the Applicant manages risk, by identifying major technical R&D

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

risks and clearly proposes feasible, effective mitigation strategies" (ARPA-E, 2015). They also explain, "ARPA-E will provide support at the highest funding level only for applications with significant technology risk, aggressive timetables, and careful management and mitigation of the associated risks" (ARPA-E, 2015). From this, we learn that ARPA-E's intent is to initially accept whatever level of risk is inherent to a high-impact project, and then to deploy active management techniques to mitigate that risk. In the next two sections, we will demonstrate empirical evidence of this strategy in practice.

In general, ARPA-E's public materials have less to say about uncertainty than about risk, though perhaps their discussion on risk serves as a stand-in for both concepts. Just as low risk and low uncertainty are commonly conflated, the implication of ARPA-E's acceptance of high risk may also be meant as an acceptance of high uncertainty. Regardless, it is clear from ARPA-E's self-descriptions that uncertainty is an essential feature of their work. Like DARPA before it, ARPA-E designs technical programs around specific challenges that could result in a transformational impact. ARPA-E states that their goal is to support research that "creates fundamentally new learning curves" (ARPA-E, 2015). In this pursuit, there is no clear path or roadmap to success. Uncertainty is therefore a desirable feature for ARPA-E's operations, due to the uncertain nature of transformational innovation compared to R&D projects that pursue incremental advances to existing technology.

ARPA-E's funding cycles begin with the hiring of a program director for a three-year term. (The routine turnover of program directors is an interesting byproduct of ARPA-E's active program management strategy, which will be explored in more depth in Section 7.) 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 not confined to their own specific area of research, and they are expected to engage with their target research communities and host stakeholder gatherings to refine their idea.[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 FOA authored by the program director. Example FOAs include Batteries for

[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.

Electrical Energy Storage in Transportation (BEEST) issued in 2010 or Full-Spectrum Optimized Conversion and Utilization of Sunlight (FOCUS) issued in 2013.

The program director then oversees a process of solicitation and merit review, which will be described in detail in Section 4.1. Ultimately, the program director is responsible for recommending to ARPA-E's Director which proposals should be selected for funding. The fact that ARPA-E employs individual decision-making in its proposal selection process means that there is the potential for greater uncertainty in their research portfolio, as described above. An increase in uncertainty could benefit the agency as they pursue transformational ideas and new learning curves. An additional benefit comes from the fact that program directors are exploring new intellectual areas; this allows them to be more creative and more accepting of novel ideas.

During the course of a project, ARPA-E program directors remain closely engaged with researchers, receiving quarterly progress updates and giving feedback on the same schedule. Project management will be discussed in detail in Section 4.2. Importantly for this work, program directors have decision-making power over the terms of a project, both in the initial negotiations and throughout the course of the project. Indeed, we find that the options described by Huchzermeier and Loch (2001), such as abandon and expand, are commonly used by program directors at ARPA-E. This gives the agency an opportunity to mitigate risk, which is clearly a component of their strategy as outlined in the FOAs.

Through the combination of practices described here, we see that ARPA-E has embraced active program management. ARPA-E provides great flexibility and autonomy to its program directors in the stages of program design, selection and project management. Furthermore, we've shown that these strategies give ARPA-E more options in the level of uncertainty and risk in its research portfolio. Using the data and results presented in the coming sections, we will show how program directors use their flexibility in practice to cope with risk and uncertainty.

4. Proprietary Data Set

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 additional data: intellectual property and market engagement outcomes (collected by ARPA-E), publication outcomes (collected by the authors from Web of Science), and company founding year (collected by the authors from public information). These data were completely scrubbed of identifying information in order to protect the confidentiality of the applicants.

In this section, we offer a brief description of the proposal and project datasets, as well as the output data collected. Some of the key variables are listed in Table 1. Selected descriptive statistics are highlighted in this section; complete summary statistics are included in Appendix B.

Table 1: List of key variables

Project Modifications	Internal Project Status Variables	External Output Variables
Net Budget Increased Net Budget Decreased Net Project Extended Net Project Shortened	Percent Quarters Red Percent Quarters Green	At Least 1 Publication At Least 1 Patent Application Market Engagement

4.1. Proposals

Before describing the proposals dataset, it is necessary to provide some background information on the proposal review process at ARPA-E. Each FOA at ARPA-E is accompanied by a Merit Review Plan, which is executed by a Merit Review Board composed of ARPA-E staff members. This board is typically chaired by the program director that crafted the program. The following summary of the proposal and selection process is based on a Merit Review Plan provided by ARPA-E, as well as discussions with ARPA-E staff.

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 expert review of concept papers, and then a subset of applicants is 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 external reviewers, who provide numerical scores and comments. Applicants are then provided with reviewer comments and are given the chance to briefly reply to these comments. 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.

At the end of the review process, the PD submits a recommendation to the Director of ARPA-E of which proposals to select.[4] 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. This oversight, in addition to the use of external review input, mitigates possible concerns over conflicts of interest in the program director's decision-making.

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:[5]

1. Impact on ARPA-E Mission Area
2. Overall Scientific and Technical Merit
3. Qualifications, Experience and Capabilities
4. Sound Management Plan[6]

We used the weights stated in the FOA for each category to calculate an overall score for each proposal-reviewer pair.[7] In total, we have 1,800 proposals for which we can compare scores and selection outcomes.

Comparing the mean overall scores for proposals that were selected to those that were not selected, it is clear that PDs use significant discretion in selecting proposals. Figure 2 illustrates the distribution of mean overall scores for funded and unfunded proposals. Rather than being divided across a sharp score cutoff, the two groups overlap significantly.

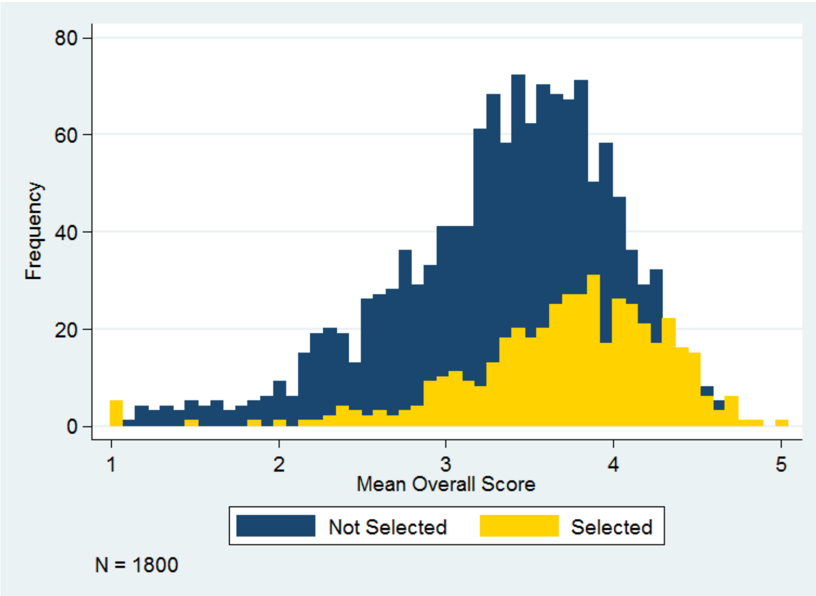
Figure 2: Mean Overall Score Distributions of Funded and Unfunded Proposals

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

[5] 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.

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

[7] FOAs for the earliest programs did not quantify the category weights. The FOA for OPEN 2009 stated that the categories as listed are "in descending order of importance," and the FOAs for Electrofuels, BEEST, and IMPACCT made no statements on the subject. Later programs in 2010 stated that the categories are of "equal weight," so we gave 25% weight to each category in those four FOAs where no percentage was stated explicitly.



The data shown in Figure 2 represent scores across 33 different technical programs. The conditions for program directors' decision-making varied significantly between programs—for example, 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 need to look at the level of the technical program.

We create an indicator variable for whether a proposal was selected based on significant discretion, as opposed to selection criteria based on a simple ranking of average reviewer scores within that program. We call these proposals "promoted," in the sense that they were selected despite a relatively low score. Our general method for identifying "promoted" proposals is to create a counter-factual score cutoff for each program; this is the cutoff that would be used if selection were based on ranking average review scores. Proposals selected from scores below this cutoff are considered "promoted."[8]

[8] We create two such cutoffs for each program: (i) based on the number of projects selected and (ii) based on the proposed budget for the FOA. In the first method, we consider the number of proposals selected under a given FOA to be N. We then place the cutoff at the Nth highest mean overall score. In the second method, we tally the cumulative proposed budgets of the proposals to a given FOA, starting from the highest mean overall score, and we compare this tally to the total budget listed in the FOA. When the cumulative budget reaches the total budget for the FOA, we place the cutoff at that score.

For nearly every program, the first method produced the lower of the two cutoffs. We use the lower cutoff to obtain a conservative estimate of the number of projects that would not have been selected without individual discretion. Proposals are labeled "promoted" if they received a mean overall score below this cutoff.

Nearly half (46%) of the selected proposals from focused FOAs are "promoted." The "promoted" variable does not apply to open programs, because the proposals span a wide range of technology types and the projects are not directly compared to each other. In section 5, we analyze the differences between projects that were "promoted" and projects that would have been funded under consensus review.

4.2. 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.

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 exclude projects that were still in progress in 2016 by limiting our dataset to those that ended on or before Dec. 31, 2015. Our dataset contains 234 completed projects.

As an ARPA-E project proceeds, awardees submit quarterly reports addressing the progress made on each of the project's negotiated milestones. Milestones may be re-negotiated and revised throughout a project. Following each quarterly report from the awardee, the PD rates the project's performance along several dimensions: technical, cost, schedule, and overall. We have quarterly report data, including percent completion of each project milestone, for the quarters beginning Jan. 1, 2012 through October 1, 2015. Our quarterly status data cover quarters starting July 1, 2010 through July 1, 2015. When status data are missing, we carry the ratings forward from previous quarters.

The ratings are given as stoplight colors (red, yellow, or green). According to our discussions with ARPA-E staff, the definitions of these colors are as follows: *green* for projects that are on track and meeting milestones; *yellow* for a project that has missed milestones but can recover;

red for a project that has missed significant ("go/no-go") milestones and may not be able to recover.

It should be noted that projects regularly change hands between PDs; 70% of projects experience at least one change in PD. Because PDs are hired on short-term contracts (three years with the possibility of renewal), projects that they initiate are often ongoing when they depart the agency. These projects continue under management by a new PD, so the recorded ratings may not be from the same PD that originally recommended the project for selection.

In addition to the milestones, the budget and timeline for a project may also be re-negotiated and revised. The project's funding may be increased or decreased, and its duration may be extended or shortened. We sum changes over time to both a project's budget and its timeline, and we measure the net change. Some projects in particular are considered "terminated" by ARPA-E, though this label does not encompass all projects that experience a shortened timeline or a reduced budget.

Our analysis focuses on four indicator variables for the four possible types of net modifications that can be made to a project: increased budget, decreased budget, extended duration, and shortened duration. The most common of these modifications is an extended project, experienced by 73% of projects. The next most common is an increased budget, which was applied to 30% of projects. Shortened project and decreased budget were both relatively less common, occurring for 12% and 13% of projects respectively.

Using the quarterly status data, we create two continuous metrics for internal status over the length of the project. First, we calculate the percent of quarters in which a project received a green status rating; then we repeat this calculation for red status ratings. These two variables allow us to measure the average performance of a project over its duration. The average ARPA-E project was overall green in 47% of its active quarters and overall red 13% of the time.

4.3. Project Outputs

In order to address the effectiveness of project selection and management practices, we need quantitative indicators of research progress. However, there is a great deal of debate around appropriate research metrics. 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.

Publication data were collected for each award through Dec. 31, 2015. Briefly, we collected these data by searching Web of Science for all award or work authorization numbers for ARPA-E projects; see Goldstein (2016) for more details on the search method.

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 reported 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.[9] 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. 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.

5. Modeling Active Program Management at ARPA-E

In this section, we describe our empirical models for project selection and management at ARPA-E. We show that individual empowerment enables PDs to select projects with a wider range of scores, which in aggregate form a portfolio of projects with greater uncertainty. The

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

PDs are then given the flexibility to manage this uncertainty. We show that PDs use budget and project length modifications to adjust the investment committed by ARPA-E in response to project progress. These modifications are used to reduce risk as new information becomes available.

5.1. Modeling Project Selection

Based on the agency's own descriptions of its procedures, we understand that ARPA-E program directors use individual discretion in project selection. In order to quantify the extent of individual decision-making, we model the relationship between various features of a proposal and the selection decision. A strong, positive association between score and selection is a feature of conventional grant-making using score cutoffs, whereas a lack of association would indicate that selection decisions are made completely independently.

We use a linear probability model to predict selection based on the score data for each proposal. The regression represented by Equation 1 captures the probability of selection as a function of one or more explanatory variables (e.g. mean overall review score), along with a fixed effect for the technical program that received the proposal:

Equation 1: $Y_i = \alpha_0 + \alpha_1 X_i + \varphi_i + \varepsilon_i$

where Y_i is the binary outcome variable for whether proposal *i* was selected; X_i is the variable of interest for application 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 Appendix B, along with results that include an additional fixed effect for organization type of the applicant.

First, we ask whether ARPA-E PDs agree with the average assessment of external reviewers when choosing which proposals to fund. In other words, is the mean overall review score of a proposal associated with ARPA-E's selection decision? Table 2 shows the results. The mean score is in fact correlated with selection; specifically, we see an 18% greater probability of selection for each additional point in the mean overall score.

Table 2: Association of Mean Review Scores with Project Selection

	(1) Selected	(2) Above Hypothetical Score Cutoff	(3) Selected
Mean Overall Score	0.179***	0.429***	
	(0.030)	(0.045)	
Mean Impact Score			0.102***
			(0.019)
Mean Merit Score			0.122***
			(0.029)
Mean Qualifications Score			-0.025
			(0.026)
Mean Management Score			-0.012
			(0.015)
Observations	1800	1800	1800
R^2	0.144	0.456	0.166
Mean of Dep. Var.	0.253	0.275	0.253

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

How strong is the association between mean score and selection? To gauge the strength of the association, we model a hypothetical scenario where selection is determined entirely by the mean overall score. In the second model of Table 2, the outcome variable is whether a proposal scored higher than the cutoff for determining "promotion." The result is a 43% increase in linear probability for each additional point in score. Comparing 43% to 18%, we learn that the association between mean overall score on selection at ARPA-E is relatively modest. The association between score and selection is less than half as strong as it would be if ARPA-E used ranking of mean scores to select proposals, rather than individual discretion.

Next, we ask which specific aspects of reviews are most relevant to the selection process at ARPA-E. We break down the mean overall score into each of the four categorical scores and test their relationship with selection. Model 3 of Table 2 shows that the mean Impact and Merit categorical scores have some predictive power on selection within ARPA-E programs; an increase of 1 point in each score leads to a 10-12% increased probability of selection. We learn that to the extent to which ARPA-E PDs agree with the average opinion of the reviewers, they are agreeing on the basis of the proposal's Impact and Merit scores. Mean Qualifications and Management scores, on the other hand, have no statistically significant effect on selection.

Having established that a proposal's overall mean score correlates with selection, we move on to investigate other features of the score distribution. Specifically, we look to the width of the score distribution to represent the level of collective uncertainty surrounding a given proposal. To understand how uncertainty influences ARPA-E's selection decisions, we ask: are the program directors' decisions associated in any way with the width of the distribution of scores, i.e. the extent of reviewer disagreement?

On its own, the standard deviation of overall scores for a proposal does not correlate with selection for a given program (Table 3). When the mean overall score is included in the regression, however, we see a positive (though not strongly significant) association between selection and standard deviation of scores. In other words, ARPA-E PDs have a mild tendency to select proposals on which reviewers disagree, given the same mean overall score within a program.

Table 3: Association of Review Score Distribution with Project Selection

	(1) Selected	(2) Selected	(3) Selected	(4) Selected
Std. Dev. Overall Score	0.018	0.111*		
	(0.065)	(0.056)		
Mean Overall Score		0.231***		
		(0.022)		
Min. Overall Score			0.035	-0.014
			(0.032)	(0.036)
Max. Overall Score			0.138***	0.057**
			(0.022)	(0.023)
Med. Overall Score				0.132***
				(0.032)
Observations	1722	1722	1800	1800
R^2	0.075	0.163	0.138	0.150
Mean of Dep. Var.	0.258	0.258	0.253	0.253

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

Reviewer disagreement may not be symmetric around the mean score—there could be outlier opinions that are either extremely positive or negative. We consider whether the two bounds on the range of scores (minimum score and maximum score) differ in how they relate to the probability of selection (Model 3 of Table 3). In a model that contains only minimum score, maximum score and program fixed effect, the results indicate a negligible effect of minimum score. This suggests that ARPA-E PDs are more likely to select proposals that were highly-rated by at least one reviewer, while the presence of a single low rating does not correlate with selection at all.

When we add median score to the regression in Model 4 of Table 3, we find that the association between maximum score and selection is dampened, but still significant. We interpret this as a PD's tendency to agree with the bulk of the reviewer scores, but also an inclination to agree with lone high scores. Given two proposals in a program with the same median score, ARPA-E PDs are likely to select the one with the highest maximum score. They use their discretion to surface proposals that have at least one champion, even if there are also detractors.

5.2. Modeling Project Modification

Active involvement of the program directors in ARPA-E projects is a unique feature of its operations. In this section, we attempt to establish quantitatively how program directors use their authority to modify the terms of projects. We model the relationship between the internal status of a project and project modifications by ARPA-E PDs, to learn whether there is a trend in how program directors choose to modify projects in response to project performance.

We use a linear probability model to predict net changes to the project in either budget or duration. Again, Equation 1 captures the probability of a project change (Y_i) as a function of an explanatory variable (X_i), such as the percent of status ratings that were rated "green" overall, while controlling for the technical program that funded the project. In Appendix B, we also examine associations between project modifications and additional control variables.

We observe the budget and project length following initial negotiations, and we observe the actual budget and project length at the end of the project's execution. These net changes may be composed of multiple smaller changes throughout the course of a project; we compare the net change with our aggregated measures of project performance, i.e. percent of project-quarters that were rated as either green or red. We include both the proportion of green statuses and the proportion of red statuses in these models. These two variables are correlated, though not co-linear; the omitted variable is "Percent Quarters Yellow". This allows us to compare projects that are performing particularly well (higher frequency of green), as opposed to simply avoiding failure (lower frequency of red).

First, we study the use of budget modifications. We find that the net amount of budget adjustments correlates significantly with status ratings (Model 1 in Table 4). Specifically, a project with more green quarters has a more positive budget change, accounting for the technical program. The association with "Percent Quarters Red" is negative, but not significant.

Table 4: Association of Internal Status Ratings with Budget Changes

	(1) Net Budget Change (Million USD)	(2) Budget Increased	(3) Budget Decreased
Percent Quarters Red	-0.268	-0.035	0.418**
	(0.194)	(0.100)	(0.171)
Percent Quarters Green	0.381**	0.249**	-0.203**
	(0.169)	(0.089)	(0.092)
Observations	233	233	233
R^2	0.252	0.232	0.251
Mean of Dep. Var.	0.199	0.305	0.124

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

We then break down the effect of status on budget changes by dividing the projects that experienced budget modifications into two groups: those with a net increase and those with a net decrease. The results (Model 2 in Table 4) show that the decision to increase a project's budget correlates with being "more green", meaning that ARPA-E increases project budgets in response to strong performance. Meanwhile, the decision to cut a budget in a given program correlates with the frequency of both red and green status ratings (Model 3), meaning that ARPA-E cuts project budgets in response to missed milestones and/or a lack of strong performance.

Over the course of a project, ARPA-E invests more than just money in a project—staff time is a significant expense for the agency, and PDs are limited in the number of projects they can actively manage at a time. As a result, shortening or extending a project also represents a change in investment. We study project length changes in the same way as budget modifications above.

Our results show that ARPA-E PDs penalize a project that misses major milestones by ending the project early. The net change in duration of a project is negatively correlated with the frequency of red status ratings, and has no relationship to the frequency of green status ratings (Table 5). This finding is reinforced by results in Model 2 and Model 3. If a project is "more red", it is more likely to end earlier than anticipated, and less likely to be extended. Model 4 shows an identical effect using the variable for whether a project was labeled "Terminated."

Table 5: Association of Internal Status Ratings with Project Length Changes

	(1) Net Project Length Change (Years)	(2) Project Extended	(3) Project Shortened	(4) "Terminated"
Percent Quarters Red	-1.285***	-0.627***	0.842***	0.746***
	(0.266)	(0.126)	(0.124)	(0.107)
Percent Quarters Green	0.170	0.019	-0.025	-0.067
	(0.134)	(0.077)	(0.056)	(0.064)
Observations	233	233	233	233
R^2	0.408	0.237	0.433	0.374
Mean of Dep. Var.	0.472	0.730	0.120	0.103

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

The associations we measure between status and project length change are all stronger than those between status and project budget, indicating that status ratings are more closely related to project length changes than they are to budget adjustments. Nonetheless, both of these associations serve as evidence that ARPA-E program directors use their flexibility to react midcourse to project performance, and that they do so in an attempt to mitigate risk from projects performing poorly, while maintaining or increasing support for those projects that have performed well.

Because we know the scores given to the proposal for each project, we have the opportunity here to determine the relationship between project performance and the prospective ratings given by external reviewers. The relationship is in fact quite small, as shown in Table 6. A "promoted" project has a slightly higher proportion of red quarters on average, and no difference in the proportion of green quarters, compared to projects with high-scoring proposals.

Table 6: Association of Low Review Score with Internal Status Ratings

	(1) Percent Quarters Red	(2) Percent Quarters Green
"Promoted"	0.064**	-0.025
	(0.025)	(0.056)
Observations	165	165
R^2	0.213	0.166
Mean of Dep. Var.	0.116	0.477

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.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

From the previous result, we learn that external review scores are not a strong predictor of ARPA-E project performance. We also ask whether these scores predict the project management decisions made by ARPA-E program directors, i.e. are “promoted” projects more likely to result in any of the four net project modifications? Results of these regressions are shown in Table 7. No significant relationship is observed, suggesting that PDs do not apply project modifications more frequently to projects that originated from low-scoring proposals. Instead, they choose whether to modify each project based at least in part on status ratings, as demonstrated in Table 4 and Table 5. Even though “promoted” projects are slightly more often rated red, this association is apparently not strong enough to appear in a statistically significant relationship between “promoted” projects and project modifications.

Table 7: Association of Low Review Score with Project Modifications

	(1) Budget Increased	(2) Budget Decreased	(3) Project Extended	(4) Project Shortened
"Promoted"	-0.032	-0.012	0.014	0.063
	(0.067)	(0.061)	(0.069)	(0.056)
Observations	165	165	165	165
R^2	0.214	0.106	0.185	0.128
Mean of Dep. Var.	0.345	0.097	0.758	0.097

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

6. Modeling Short-Term Productivity of ARPA-E’s Portfolio

In the previous section, we arrived at an empirical understanding of how ARPA-E program directors’ selection decisions relate to prospective uncertainty, and how their decisions to modify terms of a project relate to their subjective performance assessments. In this section, we attempt to connect these concepts to external metrics of performance. 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. By analyzing the project-level determinants of this productivity, we can evaluate the effectiveness to-date of ARPA-E’s approach to project selection and management.

Modeling the probability of external outputs requires the inclusion of several control variables; there are several inherent features of a project that impact the rate of publishing, patenting and/or market activity. External outputs are associated with both the organization type (university, for-profit, etc.) and project size, measured by the project length and funding amount. (In Appendix B, we test the effect of including each of these control variables.)

First, we ask whether the use of individual discretion by ARPA-E PDs results in better or worse productivity; in other words, is there an association between the "promoted" variable and any of the external metrics? Here we control for the initially negotiated project size, in order to compare projects that were prospectively similar at the outset. The regressions in Table 8 take the form shown in Equation 2:

$$Y_i = \alpha_0 + \alpha_1 X_i + \ln(initial\ project\ length_i) + \ln(initial\ funding\ amount_i) + \varphi_i + \varepsilon_i$$

Table 8: Association of Low Review Score with External Metrics

	(1) At Least 1 Publication	(2) At Least 1 Patent Application	(3) Market Engagement	(4) Any External Output	(5) All External Outputs
"Promoted"	-0.051	0.025	0.040	0.044	-0.009
	(0.077)	(0.117)	(0.082)	(0.071)	(0.072)
Observations	165	165	165	165	165
R^2	0.368	0.327	0.290	0.362	0.171
Mean of Dep. Var.	0.455	0.424	0.321	0.745	0.085

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 and the log of initial project length, as well as a fixed effect for technical program and a fixed effect for the organization type.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

The results show that there is no significant trend among any of the external metrics. Projects that are "promoted" from a low review score are statistically indistinguishable in terms of output from those that received high review scores.

Next, we seek to understand whether the decisions that ARPA-E PDs make to modify a project correspond to any measurable difference in the productivity of these projects. We use four measures of project modification (extend or shorten projects, increase or decrease budgets) and test the relationships of each separately. The regressions in Table 9 take the form shown in Equation 3:

$$Y_i = \alpha_0 + \alpha_1 X_i + \ln(actual\ project\ length_i) + \ln(actual\ funding\ amount_i) + \varphi_i + \varepsilon_i$$

Here we control for the actual project size, in order to compare projects that are retrospectively similar in execution. In the previous section, when modeling proposal selection, we intended to compare proposals that were prospectively similar, i.e. the same level of agency investment requested by the applicant. However, when we model project outputs, we must take into account the inherent effect of investment size on productivity. Therefore projects must be compared on the basis of their actual investment, rather than their intent in the proposal stage. For example, if a project is awarded $10 million but is terminated after $1 million is spent, then we cannot properly understand the performance of this project by comparing it to other $10 million projects. The correct metric for productivity of this investment is its output compared with other $1 million projects.

Table 9: Association of Project Modifications with External Metrics

	(1) At Least 1 Publication	(2) At Least 1 Patent Application	(3) Market Engagement	(4) Any External Output	(5) All External Outputs
Budget Increased	0.001	0.083	0.114**	0.052	0.040
	(0.074)	(0.074)	(0.049)	(0.036)	(0.046)
Observations	233	233	233	233	233
R^2	0.348	0.323	0.279	0.419	0.151
Mean of Dep. Var.	0.429	0.438	0.330	0.742	0.094
Budget Decreased	-0.063	-0.303***	-0.122*	-0.180	-0.051**
	(0.083)	(0.104)	(0.060)	(0.118)	(0.023)
Observations	233	233	233	233	233
R^2	0.349	0.352	0.277	0.432	0.151
Mean of Dep. Var.	0.429	0.438	0.330	0.742	0.094
Project Extended	0.074	0.124	0.140	0.091	0.067
	(0.059)	(0.076)	(0.100)	(0.065)	(0.046)
Observations	233	233	233	233	233
R^2	0.350	0.327	0.282	0.422	0.155
Mean of Dep. Var.	0.429	0.438	0.330	0.742	0.094
Project Shortened	-0.137**	-0.190	-0.365***	-0.330***	-0.156*
	(0.065)	(0.140)	(0.101)	(0.113)	(0.078)
Observations	233	233	233	233	233
R^2	0.354	0.330	0.316	0.459	0.169
Mean of Dep. Var.	0.429	0.438	0.330	0.742	0.094
"Terminated"	-0.074	-0.268	-0.286**	-0.382***	-0.095
	(0.076)	(0.158)	(0.123)	(0.107)	(0.101)
Observations	233	233	233	233	233
R^2	0.349	0.339	0.297	0.470	0.156
Mean of Dep. Var.	0.429	0.438	0.330	0.742	0.094

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 actual award amount and the log of actual project length, as well as a fixed effect for technical program and a fixed effect for the organization type.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

We find that there are several significant relationships between the externally measurable outputs and two of the four types of project size adjustments (Table 9). Specifically, decreasing a project's budget and shortening its timeline both negatively correlate with multiple measures of productivity. These regressions compare a project that was shortened or experienced a budget cut, with other projects that in the end, lasted as long or spent as much money, but had not been cut. A project that experienced a budget cut is less likely to have filed a patent application or achieved some form of market engagement. A project that was cut short in duration is less likely to have published a paper or engaged the market.

The common feature between the decision to cut a budget and the decision to shorten a project is that both actions reduce the agency's investment in a project, compared to the originally negotiated award terms. We know from previous analysis that PDs cut budgets and limited project length for those projects that they perceived to be underperforming. Now, we can confirm that their attempt to cut those projects did in fact serve to mitigate the risk carried by those underperforming projects.

There are minimal relationships between external outputs and the choice to extend a project or increase its budget, both of which would serve to increase ARPA-E's investment in a project. Given our small sample size, we can't determine the size of these relationships with confidence. We also note that external outputs tend to relate more strongly to project length changes than they do to budget adjustments, as was the case for internal status ratings as well.

Taken together, the regressions shown here and in Section 5.2 demonstrate that ARPA-E program directors modify projects to mitigate risk based on their subjective assessment, and that these modifications also relate to externally-measured project performance. We can take this a step further and directly determine how closely program directors' status ratings relate to our external markers of impact. Table 10 shows that, in fact, projects that receive primarily green internal status ratings are more likely to produce a publication, patent and achieve a measure of all external outputs. On the other hand, those that receive primarily red internal status ratings are less likely to engage with the market or attain any sign of external output. This tells us that program director's subjective assessments and our external metrics are complementary in terms of measuring project performance.

Table 10: Association of Internal Status Ratings with External Metrics

	(1) At Least 1 Publication	(2) At Least 1 Patent Application	(3) Market Engagement	(4) Any External Output	(5) All External Outputs
Percent Quarters Red	-0.047	-0.153	-0.438***	-0.468***	0.023
	(0.124)	(0.189)	(0.122)	(0.148)	(0.072)
Percent Quarters Green	0.218**	0.263*	0.074	0.123	0.174**
	(0.093)	(0.138)	(0.138)	(0.098)	(0.082)
Observations	233	233	233	233	233
R^2	0.364	0.339	0.301	0.464	0.178
Mean of Dep. Var.	0.429	0.438	0.330	0.742	0.094

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 and the log of initial project length, as well as a fixed effect for technical program and a fixed effect for the organization type.
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

7. Discussion

In the previous section, we modeled elements of ARPA-E's project selection, management and impact. In this section, we discuss those findings and offer some interpretation.

7.1. Individual Empowerment in Project Selection Enhances Uncertainty at ARPA-E

Our results show both (1) that ARPA-E employs individual decision-making in project selection and (2) that they use it to accept uncertainty into its portfolio.

To the first point, nearly half of the selection decisions made by ARPA-E PDs diverge from the average opinion of external reviewers. Selection decisions do correlate with mean scores, specifically on the project's Impact and Merit, but this correlation is much lower than it would be for a consensus peer review system. In particular, selection decisions appear to be completely independent of review scores on the criteria of Qualifications and Management.

Secondly, proposals on which reviewers disagree are more likely to be selected, controlling for mean score. This indicates that the research portfolio resulting from individual decision-making at ARPA-E carries more collective uncertainty than one based on consensus peer review.

ARPA-E PDs are tolerant of low scores, and they are particularly inclined to choose proposals that received a high maximum score, controlling for median score.

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 between individuals.

Having described the average selection decisions made within an ARPA-E program, we consider the impact of these decision-making strategies on the early outcomes of the ARPA-E portfolio. We see no statistical difference between low-scoring and high-scoring projects across all the metrics of impact that we measure. At this early stage of assessment, it seems as though ARPA-E has been able to compile a portfolio of projects with greater uncertainty, which helps fulfill their mission to pursue "white space," without seeing any downside in terms of short-term productivity.

Without long-term data on the fate of ARPA-E projects, we can only speculate what will take place as these technologies continue to develop. We expect there to be some connection between early outcomes and the chance of future success. If the lower-scoring projects end up being similar in long-term outcomes, then the decision-making strategy will have had no direct impact on the productivity of ARPA-E's investment. If however, the selected projects have more divergent outcomes, due to a higher level of uncertainty *ex ante*, then individual decision-making could actually result in greater long-term impact of ARPA-E.

Indeed, divergent outcomes are what one would expect if selection is in fact skewed toward proposals with great potential for impact, as ARPA-E's self-description indicates. The ARPA-E motto, "If it works, will it matter," illustrates this idea—even if more projects were to result in technical failure, this effect could be overwhelmed by a few hugely impactful projects, resulting in overall greater long-term productivity of the agency's investments.

7.2. Individual Empowerment in Project Modification Mitigates Risk at ARPA-E

In the previous section we show that ARPA-E's portfolio of projects carries greater uncertainty than a peer-reviewed portfolio, and that the outcomes from that portfolio appear to be indistinguishable in the short term. Turning to management of ongoing projects, we gather insight into how ARPA-E is able to manage the inherent risk in its portfolio. We find that (1) ARPA-E PDs regularly take advantage of their flexibility under active program management by modifying the terms of a research project, and (2) that they do so in an attempt to mitigate the agency's level of risk.

The first point, that ARPA-E program directors have flexibility to modify the projects that they manage is evident from the large proportion of projects (88%) that are modified from their original terms. Program directors also exchange feedback regularly with awardees and have the ability to change the project's technical milestones. These milestone changes are a key component of managerial flexibility, and they occur regularly as well; we observed addition of new milestones or deletion of existing milestones in 45% of projects.

The second finding, that project modifications serve on average to mitigate perceived risk, is based on the association between internal status and net changes to the budget or timeline of a project. Though editing of milestones is an important type of modification, our data do not allow us to draw a connection between the change in milestones and the level of investment of the agency. It is logically straightforward to claim that ARPA-E's investment in a project scales with both the project budget and the duration of the project. Because risk relates to the technical failure of a project and loss of the investment, we limit our discussion to the budget and timeline modifications specifically.

We found that ARPA-E's decision to modify a project is generally sensitive to the full range of project status from red to green. Of the four modifications we measured, two (net budget increase and net budget cut) showed a relationship with frequency of green status. Three of the four (all except net budget increase) showed a relationship with frequency of red status. This indicates that both the transition to red and the transition to green influence the likelihood that a project will have its terms adjusted. Moreover, when there is a relationship between status and project modifications, the sign of the relationship is consistent—poor performance is more likely to be punished with a reduced investment, and high performance is likely to be rewarded with an increased investment.

Among the four types of project modifications, extending a project's timeline is by far the most frequently deployed. And yet, the decision to shorten a project's timeline is most strongly associated with project status. Projects are shortened in response to red ratings and show no sensitivity to green ratings. This means that PDs on average tolerate yellow ratings (some missed milestones), but are likely to choose to shorten the project when those missed milestones become significant and the project status switches to red. This serves to limit ARPA-E's exposure to risk by cutting short a project that shows signs of failing.

In the context of a high-risk, high-reward program such as ARPA-E, it is important to note that "failing" projects, as described here, do not represent a systemic failure of the organization. Projects that do not meet their technical goals indicate an organizational culture willing to take risks on long-shot projects, as mandated in the authorizing legislation for ARPA-E. In fact, having too few failed projects would be concerning for ARPA-E, considering its mission to pursue high-risk, transformational research.

It should also be noted that throughout this discussion, we assume a particular mechanism for the relationship between status and modifications: PDs propose modifications to a project based on its technical status, i.e. status has a causal effect on modifications. It is also possible that project modifications throughout the course of a project have an effect on its subsequent performance, but we assume this effect to be small relative to the opposite direction of causality, based on our discussions with PDs.

Finally, we are able to assess the effectiveness of program directors' managerial flexibility in promoting short-term project outputs. We learn that shortened projects are in fact less likely to have yielded any measurable success, relative to those that were initially contracted to be a shorter length. The same is true for those that experience budget cuts. PDs choose to reduce investment on less productive projects, and the short-term productivity of the agency is enhanced as a result, by freeing up resources from those projects.

It should be noted that we assumed in constructing the model that we could measure project performance at the time of the program director's decision to modify the project, using both the internal status and external output variables. This assumption is only true in the case if the project ends shortly after the terms are modified, which is likely the case for shortened projects and decreased budgets. For extended projects and increased budgets, the time period following the decision to expand the investment also factors into our measures of productivity, and so our measurement is less reflective of the project status that informed the program director's management decisions.

Our discussion here pertains directly to the idea presented in Section 2.2 of two opposing strategies of project management: freedom for researchers or flexibility for managers. On first glance, these strategies appear directly contradictory, and so there must be a conflict between those who find one or the other effective in promoting innovation. Is investigator freedom *à la* HHMI the best strategy for a research program, or is it better to have managerial flexibility as they do at ARPA-E?

Our view is that it is critically important that *some qualified individual* has the flexibility to change research directions, whether that person is the investigator or program staff. Investigators at HHMI have the freedom to modify their research plans in response to new information as research progresses. Program directors at ARPA-E have the freedom to modify the milestones and terms of the research project in the same way. Research is inherently uncertain, and project aims should be expected to change. If the program staff members are well-informed on the status of the project, and well-versed in the technical field, the fact that they have decision-making power may carry some of the same benefits as if the investigator had that power.

Applying the findings of Manso (2011) to active program management at ARPA-E helps illustrate this idea. In Manso's model, a researcher is contracted to deliver a particular outcome, and they must choose an action when the work begins. In this setting, threat of termination incentivizes use of known actions, as opposed to innovating and exploring the unknown. In the case of ARPA-E, however, the researcher crafts their proposal around certain actions before work begins. The program director, taking advantage of their considerable freedom to explore, designs the technical program to be innovative and inherently exploratory. As a result, exploiting known actions is not an option, and the threat of termination makes ARPA-E awardees more likely to adhere to the negotiated work plan, rather than diverting effort to other plans.

7.3. Active Program Management is Not Universally Appropriate

It is important to note that active program management as practiced by ARPA-E is not suitable for all research funding programs. ARPA-E is well-suited for these practices for three reasons: its intent to fund transformational research in technological "white space," its specific mission of advancing energy technology, and its ability to hire top experts as program directors.

First, as discussed in Section 3, uncertainty is especially important for ARPA-E as an agency that pursues “white space”; these areas are inherently less likely to be covered by a well-defined research community with an established knowledge base. And at the same time, flexibility is especially important to managing projects with great uncertainty (Dixit and Pindyck, 2004). The more uncertain the research area, the less appropriate consensus peer review would be, given its bias against novelty. We note that this finding only applies to the framework of proposal selection. If the unit of funding is the investigators themselves, as it is at HHMI, then the investigator is the individual selecting the initial research direction and they may in fact choose to pursue highly uncertain ideas.

Second, the specificity of ARPA-E’s mission requires that they give flexibility to managers and not to researchers. The agency aims to carry out R&D in the technological areas outlined by each FOA. If instead they wanted to promote scientific advancement in general, they could surely save money by reducing program director responsibilities and giving investigators freedom to continue their project as originally negotiated. As new information comes in, the ability of the project to support ARPA-E’s mission (in other words, the likelihood that the project will succeed in terms of its original goals) may change. ARPA-E program directors are best able to judge the value of the project to the agency’s mission, both prospectively and on an ongoing basis.

Finally, the importance of hiring the highest quality staff to carry out active program management merits discussion. A program where individuals have authority to make decisions is completely dependent on the quality of each individual’s judgment and ability to make smart decisions. Therefore, it is essential that these individuals be top caliber experts in their technical field. In terms of project selection, the program must ensure that an individual staff member is as knowledgeable as possible, such that their own personal uncertainty is as close as possible to the collective uncertainty of the community of experts. The choice to select uncertain projects should reflect the limits of the field as a whole, rather than a single individual’s lack of knowledge. The quality of the individual staff member’s judgment is also critical to the successful implementation of managerial flexibility. If program directors exercise authority to influence the direction of a project, they can contribute significant benefits or significant harm, depending on their technical expertise and their management skills.

Based on our conversations with ARPA-E program directors, we believe ARPA-E's hiring practices are suited especially to the recruitment of high quality experts. First, the fact of empowerment itself attracts applicants who are motivated to make an impact in the agency's mission areas. This implies that changing the job description of existing program staff to include empowered decision-making would not be as effective. Second, the short-term contract (three years with the possibility of renewal) attracts top-tier talent from both academia and industry, who continue on their career path after a stint at ARPA-E.

One side effect of short-term hiring is that when the program director's contract expires, the projects they manage must transition to a new program director. If, as we argue, short-term hiring is essential to attracting talent, and if staff talent is central to the success of active program management, then these project hand-offs are a necessary aspect of active program management at ARPA-E. We have found that a majority (70%) of projects experience a change in program director, and that these projects are not statistically different in terms of performance (analysis not shown here), so there is no evidence of harm from short-term hiring.

8. Conclusion

Unlike other grant-making organizations within the federal government, ARPA-E has internalized active management principles into its program design to empower its program directors. Program directors at ARPA-E are empowered in two important ways: first in the project selection phase and then in the project management phase. In the first phase, PDs are empowered to use their own subject matter expertise to select proposals for funding, and in doing so, they create a portfolio of projects that retains, rather than eschews, uncertainty. In the second phase, the PDs are empowered to use their own judgment to modify the terms of the project, which they do to mitigate risk exposure. ARPA-E staff are able to use greater flexibility in selection and management to create a portfolio or projects that contains the possibility for great impact, while managing downside outcomes to minimize loss.

Through a quantitative analysis of the data on ARPA-E project selection and management, we have described how ARPA-E PDs act on these freedoms, and we have measured the effects on the productivity of the agency's research portfolio. PDs use their freedom to select proposals that would not have been selected based on external review scores, and therefore would likely have had less success seeking funding elsewhere. During the execution of these projects, PDs use their freedom to cut projects short if they are progressing poorly, or otherwise tailor the level of agency investment according to the project's technical status. The sum of these two types of empowerment is a two-pronged strategy: ARPA-E accepts prospective technical uncertainty upfront, and then reduces support for failing projects during execution.

It is worth re-iterating that the fact of program directors' empowerment at ARPA-E does not itself imply any particular outcome for the agency, either in risk, uncertainty, or project performance. Empowered PDs could instead choose projects that overlap exactly with the consensus opinion of external reviewers, thus limiting the amount of uncertainty the agency takes on board. They could choose to extend projects rather than cut them in response to missed milestones, thus tolerating high levels of risk. In this paper, we've provided evidence of how ARPA-E PDs use their freedoms in practice. These findings point to an intentional choice on the part of the agency, in hiring PDs or training them or both, to encourage high impact projects to try and fail fast.

9. References

coming soon

10. Appendix A: Descriptive Data

10.1. Proposals

The dataset of full proposals to ARPA-E contains 2,335 proposals submitted through Dec. 31, 2015. Proposals in our datasets were submitted in response to FOAs for 36 different technical programs.[10] A majority of these programs were targeted toward particular technology areas; we call these "focused" programs, as opposed to "open" programs where applicants submitted proposals across all areas of advanced energy technology. 24% of full proposals were selected for negotiation.[11] The selection rate for full proposals varied widely by program.

[10] Five programs (IDEAS, GENSETS, MOSAIC, SHIELD, and SWITCHES) had two FOAs. The FOA for the CHARGES program was not publicly available and was only distributed to eligible applicants by invitation.

[11] Some projects marked as "Selected" did not become projects, and some projects marked as "Not Selected" did in fact become projects. Our variable for selection includes both of these categories.

We have statistics on external review scores for 1,800 of the 2,335 full proposals. The number of reviews per proposal ranges from 1 to 18, with a median of 3.[12] The mean overall scores for these proposals cover the full range of scores from 1 to 5, with an average of 3.4. The standard deviation of overall score for the average project is 0.68, though it ranges from 0 to 2.6. For 20% of proposals, the standard deviation of overall scores is greater than 1 point. This suggests that there is a significant extent of reviewer disagreement regarding the quality of some proposals.

Comparing the mean overall scores among selected proposals to those that were not selected (Figure 2), we see that the selected proposals are not differentiated by their scores. The mean overall score for selected projects was 3.7 out of 5, compared to 3.3 out of 5 for non-selected projects. By ranking the proposals in order of mean overall score within a program, we see that PDs frequently choose proposals from across the full range of scores.

Nearly half (49%) of all the selected proposals in the dataset are marked as "promoted." The proportion of "promoted" proposals ranges from 10% to 91%, suggesting that PD discretion is not applied uniformly across programs.

10.2. Projects

Our projects dataset includes 471 projects initiated by the end of 2015. Because we limit our projects dataset to only completed projects, we exclude those that were active at the start of 2016. This leaves 234 projects with end dates on or before Dec. 31, 2015. The median start date for a project in our dataset is Dec. 9, 2011, the median end date is July 22, 2014. The dataset includes 25 projects (11% of all projects) that were marked as "terminated." The average project length for a "terminated" project is 1.7 years (approximately 7 quarters) compared to a mean length for all projects of 2.6 years.

Twenty-three programs to date have yielded at least one completed project for inclusion in this dataset. The projects are distributed among five organization types (universities, established companies, startup companies, and non-profits), where startup companies are defined as being founded within 5 years before the start date of the project. The largest number of lead recipients

[12] Proposals with only one review were excluded from the regressions that include standard deviation of scores as an explanatory variable.

was from universities. Projects vary widely in their initially-negotiated funding amount; the mean award at the start of the project is $2.3 million USD, and the largest is $9.1 million USD. Projects also vary in their originally negotiated time period, from two quarters up to 4 years.

Budgets and timelines are modified mid-project for many of the projects in the dataset. The average project is extended by 0.5 years, and has a budget increase of $0.2 million USD. A majority of projects (85%) have some change in project length, and 43% have some change in budget; only 12% of projects are completed with their original budget and duration. The frequency of these award modifications is evidence of active management by ARPA-E PDs.

There is a variable in our data indicating projects that are "Terminated"; these projects are labeled as "Cancelled" on ARPA-E's website. We test this variable separately from whether a project was shortened or experienced a budget cut, because they do not overlap perfectly. The correlation coefficient between "Terminated" and "Project Shortened" is 77%; for "Terminated" and "Budget Decreased", it is 41%.

The frequency of milestone changes demonstrates the dynamic nature of ARPA-E projects. With this data set, we are limited to measuring only the subset of milestone changes which involve creating a new milestone or deleting an old one. A large portion of projects (45%) experience a milestone change of this type.[13]

There are four types of status ratings given to a project each quarter: three component ratings (cost, schedule, technical) and an overall rating. Within the dataset of project-quarters, the correlation coefficients of the component status ratings with the overall status ratings (0.34, 0.71, and 0.86 respectively) indicate that "overall" is most reflective of a project's technical status. A relatively small proportion of project-quarters (7%) are overall red; 38% are yellow and 55% are green.

61% of projects only have overall yellow and green status ratings (never red), and 9% of projects only have overall green status ratings. Among the projects that do have overall red status ratings, the most common number of consecutive red quarters is 2; this means that after two consecutive quarters of red status, projects tend to either terminate or change status to green or yellow.

[13] Our record from ARPA-E's project management software begins in January 2012, so we are only able to partially measure milestone changes for projects that were in progress at that time.

Of the external metrics measured, publications and patent applications are both common outcomes, achieved by 43% and 44% of the projects in our data respectively. 33% of projects have some indicator of market engagement. 74% of projects achieved “Any External Output”, and only 9% achieved “All External Outputs.”

11. Appendix B: Additional Analysis

We test the inclusion of various controls in the model for estimating the association between mean overall score and selection; this relationship is not especially sensitive to the inclusion of control variables. The greatest predictive power (R^2 = 0.14) in Table B1 is from Model 4, which includes only a fixed effect for the technical program. This is the specification used in Table 2 and Table B2 as well. The proposed budget and organization type do not correlate strongly with selection, though there is a slightly higher tendency for proposals with a larger budget to be selected, given the same mean review score.

Table B1: Control Variable Testing for Predicting Project Selection

	(1) Selected	(2) Selected	(3) Selected	(4) Selected	(5) Selected
Mean Overall Score	0.160***	0.169***	0.176***	0.179***	0.179***
	(0.030)	(0.034)	(0.036)	(0.030)	(0.038)
Proposed Budget (Million USD)		0.014**			0.018***
		(0.006)			(0.007)
Org. Type Fixed Effect	N	N	Y	N	Y
Program Fixed Effect	N	N	N	Y	Y
Observations	1800	1463	1460	1800	1460
R^2	0.061	0.070	0.067	0.144	0.134
Mean of Dep. Var.	0.253	0.284	0.284	0.253	0.284

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

The estimated correlations between score and selection presented in Table 2 and Table 3 are robust to specification using logit rather than OLS (Table B2 and Table B3). The values listed in this table and all following logit models are odds ratios for the outcome, corresponding to a one-unit increase in the independent variable. An odds ratio <1 indicates a negative correlation. The odds ratio for Mean Overall Score in Model 2 of Table B2 is effectively infinite, because the dependent variable is a step function determined by the ranking of review score within a given program.

Table B2: Logit Specification – Association of Mean Review Scores with Project Selection

	(1) Selected	(2) Above Hypothetical Score Cutoff	(3) Selected
Mean Overall Score	3.709***	3.32e+12**	
	(1.093)	(3.79e+13)	
Mean Impact Score			2.058***
			(0.363)
Mean Merit Score			2.162***
			(0.422)
Mean Qualifications Score			0.836
			(0.171)
Mean Management Score			0.991
			(0.143)
Observations	1800	1800	1800
Pseudo R^2	0.141	0.864	0.160
Mean of Dep. Var.	0.253	0.275	0.253

Notes: Exponentiated coefficients. Standard errors in parentheses. All regressions are logit 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 B3: Logit Specification – Association of Review Score Distribution with Project Selection

	(1) Selected	(2) Selected	(3) Selected	(4) Selected
Std. Dev. Overall Score	1.108	2.645***		
	(0.394)	(0.972)		
Mean Overall Score		6.247***		
		(1.430)		
Min. Overall Score			1.237	0.935
			(0.233)	(0.174)
Max. Overall Score			3.350***	1.992***
			(0.630)	(0.423)
Med. Overall Score				2.350***
				(0.425)
Observations	1722	1722	1800	1800
Pseudo R^2	0.064	0.165	0.140	0.153
Mean of Dep. Var.	0.258	0.258	0.253	0.253

Notes: Exponentiated coefficients. Standard errors in parentheses. All regressions are logit with robust standard error, clustered by technical program. The models include a fixed effect for technical program.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Next, we test the inclusion of various controls in our estimation of the association between internal status ratings and project modification decisions. Of the four project modifications, two examples are shown here: budget decreased (Table B4) and project shortened (Table B5). These are the modifications with the strongest association with internal status ratings. Both variables show a robust relationship with proportion of red quarters. In both cases the model with the greatest predictive power includes the technical program fixed effect, which is used for the specifications in Table 4 and Table 5.

Table B4: Control Variable Testing for Predicting a Net Budget Decrease

	(1) Budget Decreased	(2) Budget Decreased	(3) Budget Decreased	(4) Budget Decreased	(5) Budget Decreased
Percent Quarters Red	0.436**	0.443**	0.404**	0.418**	0.425**
	(0.159)	(0.160)	(0.158)	(0.171)	(0.173)
Percent Quarters Green	-0.155	-0.158	-0.181*	-0.203**	-0.213**
	(0.091)	(0.093)	(0.100)	(0.092)	(0.094)
Ln (Init. Project Length)		-0.054			-0.050
		(0.053)			(0.072)
Ln (Init. Award Amount)		0.012			-0.025
		(0.032)			(0.057)
Org. Type Fixed Effect	N	N	Y	N	Y
Program Fixed Effect	N	N	N	Y	Y
Observations	233	233	233	233	233
R^2	0.162	0.165	0.185	0.251	0.274
Mean of Dep. Var.	0.124	0.124	0.124	0.124	0.124

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 B5: Control Variable Testing for Predicting a Net Shortened Project

	(1) Project Shortened	(2) Project Shortened	(3) Project Shortened	(4) Project Shortened	(5) Project Shortened
Percent Quarters Red	0.858***	0.840***	0.836***	0.842***	0.778***
	(0.109)	(0.105)	(0.109)	(0.124)	(0.140)
Percent Quarters Green	-0.006	0.001	-0.017	-0.025	-0.021
	(0.050)	(0.050)	(0.049)	(0.056)	(0.055)
Ln (Init. Project Length)		0.124*			0.252*
		(0.066)			(0.128)

Ln (Init. Award Amount)		-0.022			-0.031
		(0.021)			(0.033)
Org. Type Fixed Effect	N	N	Y	N	Y
Program Fixed Effect	N	N	N	Y	Y
Observations	233	233	233	233	233
R^2	0.375	0.393	0.405	0.433	0.492
Mean of Dep. Var.	0.120	0.120	0.120	0.120	0.120

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

The estimated coefficients for binary outcome variables in Table 4 and Table 5 are also largely robust to specification using a logit model, with one exception. The odds ratio for whether a project will have a net budget decrease given a different proportion of red status ratings (Model 2 in Table B5) is not significant. Instead, the only significant predictor of a decreased budget is the proportion of green status ratings; a project with more green quarters is less likely to have a net budget cut. Note that there are fewer observations included with the logit model than using OLS, because projects are dropped if they are in technical programs without variation in the quantity of interest (no decreased budgets or all decreased budgets).

Table B6: Logit Specification – Association of Internal Status Ratings with Project Modifications

	(1) Budget Increased	(2) Budget Decreased	(3) Project Extended	(4) Project Shortened	(5) "Terminated"
Percent Quarters Red	0.672	6.391	0.038***	1610.681***	111.696***
	(0.562)	(7.401)	(0.032)	(2667.313)	(161.475)
Percent Quarters Green	4.329***	0.025***	1.182	0.265	0.053
	(2.271)	(0.031)	(0.641)	(0.416)	(0.109)
Observations	211	165	213	173	165
R^2	0.142	0.259	0.179	0.467	0.392
Mean of Dep. Var.	0.318	0.176	0.704	0.162	0.145

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

We then test the correlations with external metrics for their sensitivity to control variables, using one example relationship each from Table 8 and Table 9. No matter what control variable structure we use (Table B7), there is no significant association between projects that were "promoted" from a low review score and those that achieve at least one external output. Meanwhile, inclusion of control variables does not change the negative association between shortened projects and achieving at least one external output (Table B8). In other words, both of these example results are robust to using different combinations of control variables.

Table B7: Control Variable Testing for Predicting Any External Output by "Promotion"

	(1) Any External Output	(2) Any External Output	(3) Any External Output	(4) Any External Output	(5) Any External Output
"Promoted"	0.071	0.108	0.018	0.084	0.044
	(0.074)	(0.071)	(0.067)	(0.071)	(0.071)
Ln (Init. Project Length)		0.101			-0.034
		(0.109)			(0.130)
Ln (Init. Award Amount)		0.127**			0.042
		(0.051)			(0.040)
Org. Type Fixed Effect	N	N	Y	N	Y
Program Fixed Effect	N	N	N	Y	Y
Observations	165	165	165	165	165
R^2	0.007	0.088	0.072	0.325	0.362
Mean of Dep. Var.	0.745	0.745	0.745	0.745	0.745

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 B8: Control Variable Testing for Predicting Any External Output by a Net Shortened Project

	(1) Any External Output	(2) Any External Output	(3) Any External Output	(4) Any External Output	(5) Any External Output
Project Shortened	-0.398***	-0.247**	-0.392***	-0.369***	-0.330***
	(0.100)	(0.097)	(0.101)	(0.116)	(0.113)
Ln (Project Length)		0.286***			0.089
		(0.091)			(0.084)
Ln (Award Amount)		0.096*			0.031
		(0.049)			(0.039)
Org. Type Fixed Effect	N	N	Y	N	Y
Program Fixed Effect	N	N	N	Y	Y

Observations	165	165	165	165	165
R^2	0.007	0.088	0.072	0.325	0.362
Mean of Dep. Var.	0.745	0.745	0.745	0.745	0.745

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

Finally, the relationships in Table 8 and Table 9 are estimated using a logit model. The results shown earlier are robust to the logit specification. No significant association is seen between projects from low-scoring "promoted" proposals and projects that have external impact (Table B9). The associations between modified projects and projects with external impacts, as observed using OLS, mostly remain—with two exceptions. The negative correlation between a net budget decrease and achieving all external outputs is no longer significant; and the correlation between a shortened project and achieving all external outputs cannot be measured because there are no projects that satisfy both conditions.

Meanwhile, a few additional relationships appear to be significant under logit that were not under OLS. All associations measured in both model types support our finding that program directors tend to either increase investment in thriving projects or decrease investment in failing projects.

Table B9: Logit Specification – Association of Low Review Score with External Metrics

	(1) At Least 1 Publication	(2) At Least 1 Patent Application	(3) Market Engagement	(4) Any External Output	(5) All External Outputs
"Promoted"	0.787	1.186	1.409	1.382	1.663
	(0.356)	(0.761)	(0.696)	(0.782)	(2.303)
Observations	155	144	151	126	80
R^2	0.283	0.196	0.209	0.168	0.164
Mean of Dep. Var.	0.484	0.486	0.325	0.746	0.175

Notes: Exponentiated coefficients. Standard errors in parentheses. All regressions are logit with robust standard error, clustered by technical program. The models include controls for the log of initial award amount and the log of initial project length, 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 B10: Logit Specification – Association of Project Modifications with External Metrics

	(1) At Least 1 Publication	(2) At Least 1 Patent Application	(3) Market Engagement	(4) Any External Output	(5) All External Outputs
Budget Increased	0.847	1.582	1.946***	1.811**	1.468
	(0.352)	(0.658)	(0.486)	(0.549)	(0.760)
Observations	216	205	212	184	138
Pseudo R^2	0.262	0.197	0.200	0.233	0.146
Mean of Dep. Var.	0.463	0.498	0.344	0.766	0.159
Budget Decreased	0.770	0.121***	0.389**	0.258	0.411
	(0.404)	(0.070)	(0.179)	(0.254)	(0.352)
Observations	216	205	212	184	138
Pseudo R^2	0.262	0.233	0.198	0.251	0.148
Mean of Dep. Var.	0.463	0.498	0.344	0.766	0.159
Project Extended	1.577	2.057*	2.434	1.932	3.396*
	(0.545)	(0.812)	(1.469)	(1.047)	(2.350)
Observations	216	205	212	184	138
Pseudo R^2	0.264	0.202	0.202	0.234	0.162
Mean of Dep. Var.	0.463	0.498	0.344	0.766	0.159
Project Shortened	0.477*	0.304	0.032**	0.129**	--
	(0.188)	(0.242)	(0.044)	(0.117)	
Observations	216	205	212	184	119
Pseudo R^2	0.266	0.207	0.249	0.280	0.147
Mean of Dep. Var.	0.463	0.498	0.344	0.766	0.185
"Terminated"	0.725	0.180**	0.106*	0.072**	0.330
	(0.309)	(0.156)	(0.123)	(0.076)	(0.505)
Observations	216	205	212	184	138
Pseudo R^2	0.262	0.218	0.222	0.300	0.152
Mean of Dep. Var.	0.463	0.498	0.344	0.766	0.159

Notes: Exponentiated coefficients. Standard errors in parentheses. All regressions are logit with robust standard error, clustered by technical program. The models include controls for the log of actual award amount and the log of actual project length, as well as a fixed effect for technical program and a fixed effect for the organization type.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$