Selling Science:

New Insights into Measuring Research Productivity

William Bianco
Department of Political Science
Indiana University

Donald J. Gerhart
Challenger Biosciences LLC

Sean Nicholson-Crotty
School of Public and Environmental Affairs
Indiana University

DRAFT: 2/12/18

Note to Workshop Participants. This work is about as early-stage as could be; in many ways it is less a presentation of findings and more a draft of a proposal for a research project based on a potentially-important insight. We are presenting it with many loose ends because are looking for advice on (a) whether the approach makes sense, (b) what needs to be changed or better-justified, (c) relevant previous work, debates, and citations, and (d) what other implications might be developed from the analysis.

Summary. We present a game-theoretic framework for assessing the intermediate productivity of high-risk, long-term programs that require a substantial up-front investment of resources as well as on-going support. Examples of such programs include not only publicly-funded endeavors led by governmental agencies (such as the construction, maintenance, and operation of the International Space Station by NASA and other national space agencies), but also privately-financed initiatives undertaken and sustained by non-governmental organizations or for-profit entities (such as the development of new therapeutic drugs and biologics by pharmaceutical firms). For such programs, periodic assessments based on *waypoints*—intermediate steps on the path toward a desired outcome—can provide significant advantages over endpoint-based assessments. Stringent assessments using waypoints that possess particular characteristics should, under certain conditions, reduce potential political interference (*sensu latu*) and increase reliance on utility-based analyses as kill-or-continue decisions are made.

This paper focuses on assessing the productivity of long-term programs that require large initial and ongoing investments of resources, but whose payoffs are uncertain and—even if eventually realized—lie far in the future. Consider the International Space Station. NASA (NASA 2015) describes the Station as a "groundbreaking research platform," with "innovative experiments" aimed at "extend[ing] results for the betterment of humanity." Yet even this glowing description acknowledges "we may not know what will be the most important discovery from the space station." At the same time, ISS operations consume well over 2 billion dollars per year for the foreseeable future. What evidence is there to suggest that Station science is on the path to producing scientific discoveries and valuable commercial innovations?

The problem goes well beyond the space program, extending more broadly to all "Big Science" programs. The enormous sums spent on these undertakings are often justified by predictions about their contributions to knowledge, the economy, or innovation. At virtually any point in the lifecycle of a Big Science program, it is not easy to explain to the public or elected officials why the program is a compelling investment, given that the benefits will emerge only over time – if at all.[1] The cost of Big Science projects can crowd out other worthy endeavors; moreover, the existence of Big Science assets—and the sunk costs they represent—generate a perverse incentive to fund projects that utilize these assets, regardless of concerns about their methods and goals.

The need for a quantitative framework for assessing progress toward impact was underscored about a decade ago in the prospectus for the federal STAR Metrics initiative, which argued, "There is currently no data infrastructure that systematically couples science funding with

[1] Recent examples include the debate over extending ISS operations (Witze 2014), and evaluations of the NSF Division of Ocean Science's fleet of research vessels (Showstack 2015).

outcomes." (STAR 2010). While this initiative created a database of federally-funded research, there has been little progress on determining whether the data collected for each project can be used to predict the ultimate value of the research. Advocates for a Big Science investment often seek to build public support by arguing that scientific progress is a complex process, where the findings of individual research projects interact, building on each other over years or decades to generate real, identifiable benefits for society. If so, how can the public be sure that researchers are on the right track, funding the right proposals, and building the right pieces of infrastructure?

At the same time, the managers of large scientific research programs face the problem of "selling science" – convincing their political masters to allow decisions to be made on the basis of scientific merit rather than political expedience. Programs that cannot distinguish good science from bad are vulnerable to intervention by political actors trying to secure benefits for their constituents. When these interventions are successful, a disproportionate share of benefits (research projects) will flow to districts represented by legislators in positions of institutional power, rather than being awarded to maximize productivity. While early studies of pork-barrel politics assumed that all programs were equally vulnerable to these pressures, more recent work argues that some agencies are more vulnerable than others. In particular, pork-barrel pressures can be resisted by agencies that have developed expertise about policy matters, giving them a viable rationale for resisting congressional demands.

This set of problems extends into the private sector as well. Entities of many types and missions—from private philanthropic organizations seeking to maximize social impact while remaining answerable to donors; to venture capitalists deploying tranches of funding into emerging opportunities for the benefit of limited partners; to corporate leaders allocating assets across high-risk yet potentially transformative innovation initiatives for the ultimate benefit of shareholders—all face pressures that analogous to those faced by NASA's leaders as they contemplate the fate of the

ISS. In the following paragraphs, while we frame cases using the terminology for government-sponsored scientific research, we are aware that our analysis is extendable to all long-term, high-risk, slow pay-off programs requiring substantial up-front investment as well as ongoing allocation of significant resources.

In essence, we are arguing that a crucial toolset is missing: a rigorous, quantitative framework that enables program managers to distinguish—rigorously, objectively, and convincingly—promising projects from those with less potential. The key question is, what variables might provide this sort of information?

Our focus here is on how evaluators might use the intermediate outputs of scientific work-in-progress, such as publications, invention reports, or patents, to revise their judgements about the merits of a research project. We refer to these measures as *waypoints— intermediate steps on the path toward a desired outcome.* To identify the characteristics that might make certain waypoints particularly valuable, this paper develops a game-theoretic model of research assessment with two players, a principal investigator (PI) who wants to secure research funding for a project, and a program officer (PO) who wishes to maximizes the productivity of a portfolio of projects under their control. PO must decide whether to fund (renew) PI's project or fund an alternative project. However, before PO makes their funding decision, they can choose a waypoint, observe whether or PI's project has achieved this waypoint, and use this signal to update their beliefs about the project's worthiness using Bayes' rule. The goal of this analysis is to identify the characteristics of waypoints that make them useful tools for assessment and to assess the tradeoff between waypoints that minimize the rate of false positives and those that minimize the probability of false negatives.

Measuring Scientific Innovation:

The Problem of Assessment

Ideally, measurements of scientific productivity should capture *endpoints* such as the creation of widely-cited knowledge or the development of new technologies and other innovations. Indeed, endpoints such as the development of new therapies are how NIH describes their "success stories," many of which involve 1980s-era grants (NIH 2017). However, as this example illustrates, considerable time (decades) typically elapse between initial funding and the realization of the desired outcomes of basic research. Thus, even if a proposal is judged to be the most effective use of resources at the time it was initially approved, without some mechanism for mid-term assessment, funds will continue to be committed for a substantial amount of time without any assurance that research is proceeding toward the intended goal. At the same time, newly-appearing proposals may be denied funds because of prior commitments to ongoing research. Finally, in the case of research programs involving multiple projects sharing goals and hardware, it may be difficult to trace endpoints back to the specific projects that generated them, or even allocate responsibility across multiple projects – meaning that final, definitive judgments on a project's efficacy may never be possible. For all of these reasons, we believe, program officers will be motivated to consider interim measures of research progress, or waypoints, as a way of refining their judgments about a project's efficacy. Assessments of scientific research face the same problem.

We describe this problem in terms of a given project X. X is one of two types: good (G) with probability $p(G) = g$ or bad ($\sim$G) with probability $(1 - p(G)) = (1 - g)$, and at the time funding decisions must be made, both PIs and POs are uncertain of Project X's type. Good projects work as advertised, yielding important scientific discoveries or valuable commercial products. Bad projects provide employment for researchers and their staff, but at best produce normal science results, filling in gaps but not generating transformational results. In general, p(G) is small, meaning only a few projects are actually good, and X's type will be observed only in the distant future.

We expect that X is one of several projects administered by a program officer PO, where the individual projects are selected to build a portfolio that maximizes the likelihood of innovations across the entire portfolio. In general, portfolio theory suggests that the optimal portfolio will consist of some number of moonshots (low probability, high variance, high value projects) and some lower value, higher probability, low variance projects. For our argument, the midterm assessment problem for the PO resolves into a set of binary decisions. Given a given of funded projects comprising a research portfolio, and a set of alternatives (available unfunded projects), should funding continue to be directed to a currently-funded project X that fills a particular niche in the portfolio, or should X be terminated in favor of an alternate project A, that would occupy the same niche in the portfolio?

f prior beliefs were the only information available, the comparison between X and A would center on the relative sizes of $p(G_X)$ *(i.e.,* the probability that Project X is good) and $p(G_A)$ (*i.e.,* the probability that Project A is good). PO would continue to fund X if $p(G_X) > p(G_A)$, and switch to A otherwise.[2] However, because X is underway, PO can revise their judgement about the merits of the proposal by considering the interim outputs of G – whether G has reached a particular waypoint (*e.g.,* a pre-determined milestone or benchmark). Just as with a medical test, the presence of a waypoint allows an observer to update beliefs about a project's promise (e.g., the probability of producing important results) based on whether the project achieved or did not achieve the waypoint.

We define the problem formally as follows. First, waypoint W takes on two values: H, meaning project G achieves or hits waypoint W, and ~H, meaning project G does not achieves or misses waypoint W.

$p(H|G)$ = probability that a good project hits waypoint W

[2] More precisely, the decision would be shaped by PO's assessment of the value of the outcomes that G and A might produce, as well as the costs of each project – we leave these complications for future work.

$p(\sim H|G) = 1 - p(H|G)$ = probability that a good project fails to hit waypoint W

$p(\sim H|\sim G)$ = probability that a bad project fails to hit waypoint W

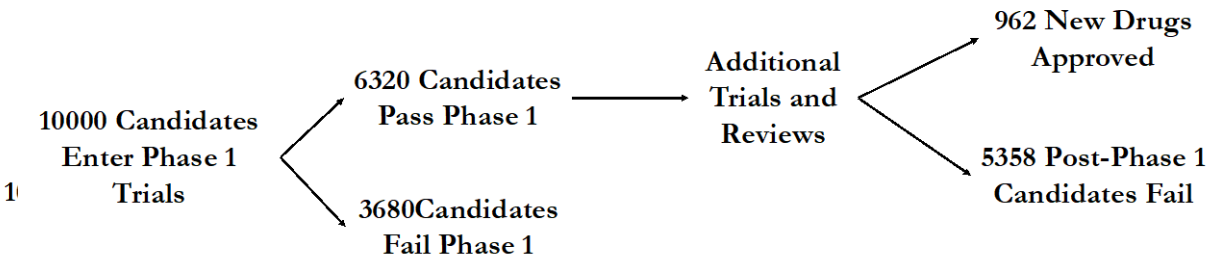
$p(H|\sim G) = 1 - p(\sim H|\sim G)$ = probability that a bad project hits waypoint W

In general, these judgments will not be perfect, meaning $p(H|G)$ and $p(\sim H|\sim G)$ will both be less than 1. A project that hits waypoint W is not necessarily good – not all projects that generate patents go on to develop important commercial technologies. Put another way, hitting a waypoint might be a false positive about the merits of Project X, and failing to hit W might be a false negative about Project X. What we expect in general is that $p(G|H)$ – which can be calculated from the two factors above plus the value of $p(G)$ – will be higher than $p(G)$ but less than 1. Conversely, suppose that a given project did not hit waypoint W. Here we expect that $p(G|\sim H)$ (again, calculable from the factors above) will be less than $p(G)$ but greater than 0. In this way, PO could use waypoints to refine their beliefs about G, which in turn increases the chances that they will make the right funding decision – funding whichever program, G or A, that has the best changes of producing transformational results.

The use of waypoints or milestones is common in academia and industry settings. In academia, for example, tenure committees assess a candidate's potential based on productivity during a relatively short probationary period. And in the case of drug development, before the first dosing of humans is undertaken, the toxicity of a candidate drug is assessed in controlled studies using non-human animals. Such assessments serve as important hurdles in every drug discovery program, since the minimum dose causing toxicity must not lie close to the dose required to deliver a pharmacologically effective level of the drug at the targeted site of action. If tests show that toxicity is problematic (*i.e.*, effectiveness cannot be achieved without excessive harm to the organism), the likelihood that the drug will ultimately be approved declines considerably and the project is almost certain to be terminated.

As a further example, consider the progress of drug candidates through the later stages of development involving tests on humans. In order for a drug to be approved by the United States Food and Drug Administration, the corporate sponsor must demonstrate, through well-designed clinical trials, that the candidate drug possesses an acceptable safety profile and delivers clinically and statistically significant therapeutic benefits to patients. Phase I studies focus on safety and involve relatively small numbers (dozens) of patients or normal volunteers. Subsequent Phase II studies provide further evidence of safety and yield preliminary insights into efficacy. Phase III (late-stage) clinical trials are complex and time-consuming, often involving large numbers (hundreds or thousands) of patients. Thomas et al. (2016) estimated that, on average, 63.2% of Phase I drug candidates succeed in Phase I. Thomas et al. further estimated that only 9.6% of Phase I candidate go to receive eventual FDA approval. For drug candidates that have completed Phase I and entered Phase II, however, Thomas et al. estimated the likelihood of eventual FDA approval at 15.3%. In this case, Phase I testing serves as a waypoint, as illustrated in Figure 1 for a hypothetical cohort of 10,000 drug candidates.

Figure 1.
Waypoints and Clinical Trials



G = Candidate will lead to approved drug
H = Candidate passes Phase 1
p(G) = .096 (prior to Phase 1)

p(H | G) = 1.0 (assumed)
p(~H | G) = 0 (assumed)
p(H | ~G) = 5358/6320 = .85

p(G | H) = .15

p(G | H) = .15

As the figure shows, 10,000 candidate drugs entered clinical (Phase 1) trials. Of these, 6320 passed Phase 1, meaning that serious side effects were not observed, while 3680 did not. Of the 6320 successful candidates, 962 survived additional trials and the FDA's review process to become approved drugs, while 5358 failed at one of these additional steps. Looking at candidates entering Phase 1, p(G), or the likelihood that any one candidate would ultimately be approved, is .096. However, using the observed percentages and assuming that all good candidates are ultimately approved, the conditional probability p(G | H), or the probability that a successful Phase 1 candidate will be approved, is .15. Put another way, passing Phase 1 provides some information about a drug candidate's potential for success (the probability of approval goes from .096 to .15, or about 56%). In contrast, failing Phase 1 yields a very precise judgment about a drug's potential, as none of these candidates were ultimately approved.

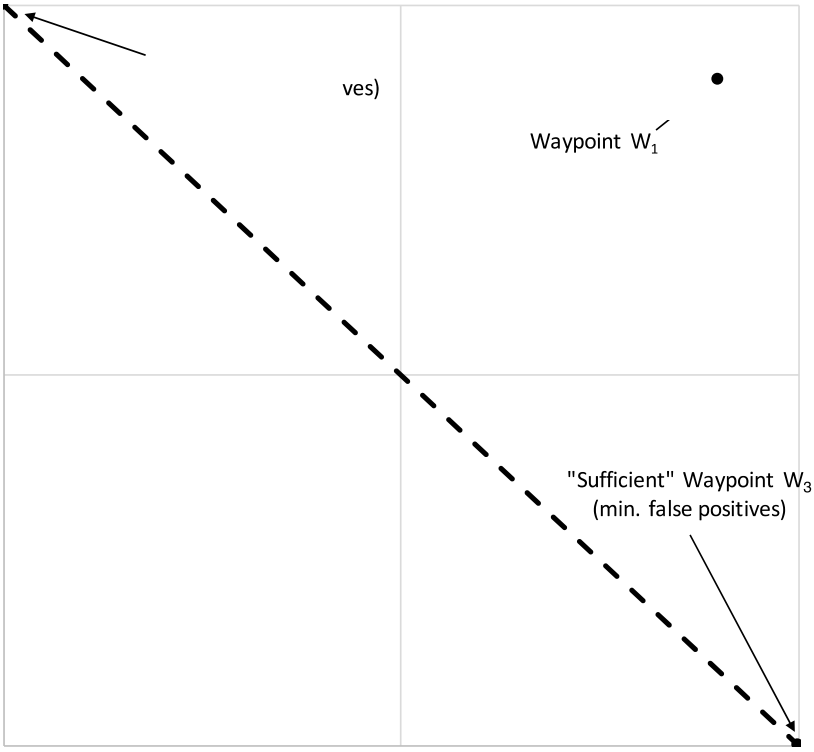
Choosing the Right Waypoint

While mid-term assessment of scientific research may be attractive, implementation is far from straightforward. Most fundamentally, mid-term judgments will generally have some element of uncertainty, as final results have yet to take shape. Thus, if achieving a waypoint is read as an indication that a project looks good, there is some chance that this judgment is a false positive, and that the project will not achieve its goals. Conversely, if failing to achieve a waypoint is read as an indication that a project looks bad, there is some chance that this judgment is a false negative, meaning that continued effort will eventually lead to project success.

Ideally, midterm evaluations would utilize a waypoint that provided definitive, risk-free information about a project's merits. Through the use of such waypoints, a PO could rest assured that any project which achieved waypoint W by time t would be destined for success, while projects failing to achieve waypoint W in a timely fashion would be bound to fail. We suspect, however, that most conceivable waypoints far short of this ideal. More than that, some waypoints might do a better job of diagnosing project success rather than project failure, or vice versa. Thus, when selecting how to judge a particular proposal, the choice is not between good waypoints and bad waypoints – that is, a waypoint that has a low probability of false positives or false negatives, versus one where these probabilities are both higher. Rather, the choice may be between waypoints that minimize false positives versus those that minimize false negatives. Cast in those terms, it is not obvious which evaluation metric is preferable.

Figure two illustrates this assessment problem. It depicts several waypoints in terms of the probability that they yield an accurate judgement given a proposal hits its waypoint (p(H|G, vertical axis) and the same probability given a negative judgment (p(~H|~G), horizontal axis).

Figure 2.
Variation in Waypoint Characteristics



We label W_2 and W_3 as "necessary" and "sufficient" waypoints, respectively – distinguishing them based on the content of their signal about G. Given W_2's characteristics (p(H|G = 1, p(~H|~G) = .5), if PO observes ~H, she knows that G cannot be good, as a good G always signals H. Thus, observing H is a necessary condition for G being a good project. At the other end, W_3 is a sufficient waypoint, meaning that if PO observes ~H, she knows that G must be a bad project.

Ideally, a PO would like to use a waypoint that has the characteristics of the waypoint W_1 in Figure 2 – good projects have a .95 chance of hitting waypoint W, and bad projects have a .95 chance of failing to do so. At least in theory, waypoints like W_1 correspond to a canonical independent external peer review process, where a panel of experts assess the merits of a project. However, studies of peer review show that the process tends to favor scientists and labs with strong reputations for producing excellent work, and disadvantages projects that use innovative methods or which ask paradigm-shattering questions. (Both become even more likely when funding constraints reduce the number of proposals that can be funded.) Thus, even a well-run peer review process will yield some false positives and some false negatives.

However, what if a waypoint with the characteristics of peer review is too costly or otherwise unavailable, and PO's choice is between a group of less-accurate waypoints, such as those corresponding to the dotted line in Figure 2? Suppose PO does not have a benchmark or waypoint corresponding to W_1, but instead has a set of potential benchmarks arrayed among the dotted line between W_2 and W_3. At one extreme, waypoint W_2 minimizes the chances of false negatives, while at the other, waypoint W_3 minimizes the chances of false positives. Assuming PO wants to maximize the probability that good proposals are funded (leading to maximizing the number of innovations from their portfolio), what is the optimal waypoint to use, or milestone to enforce, as part of their administration of project G?

As an example of the tradeoff depicted in Figure 2, consider waypoints that track publications in refereed journals. Imagine that project X is some years into the research process – too early to expect its ultimate results, but far enough along that initial findings should have been produced. For project X, publication of early results in a top-tier journal provides a strong signal of the project's promise, while failure to do so provides very little information: given the vagaries of the review process, even noteworthy early results might fail to gain enough referee support for

acceptance in a top-tier journal. Thus, if this waypoint is used to assess project X, the likelihood of a false positive is relatively low, while the probability of a false negative is relatively high – corresponding to the hard waypoint W_3 in Figure 1.

In contrast, consider evaluating G using a different waypoint, one that measures whether a project has yielded publications in any refereed journal. Hitting this waypoint is not a strong signal of a project's merits. However, a failure to hit this waypoint (no publications at all) would have negative implications for a project's ultimate output. Thus, this waypoint resembles the point W_2 in figure one, meaning a relatively high likelihood of a false positive, and a relatively lower probability of a false negative.

An analysis of waypoints must also incorporate the interaction between PI and PO – specifically, the expectation that PO's use of waypoints might affect PI's actions. The simplest possible assumption is that PI's behavior is fixed, making waypoint W a clean signal of the project's worth. However, pursuit of the waypoint by PI may involve some marginal cost, such as reducing the probability of achieving a project's ultimate goals. If so, PI might opt against taking actions consistent with W, in order to focus efforts on the project's ultimate goals.

A second complication concerns PO's selection between the two projects, the one currently funded (X) and the alternative (A). Suppose PO values G and A differently. Given the right combination of probabilities (and values), PI might opt against pursuing a waypoint chosen by PO, on grounds that Project X's prior makes it attractive enough that PO will not switch funding to A, even if the waypoint is not achieved. Here again, a waypoint that is attractive in terms of its power to distinguish good projects from bad may not be a useful benchmark in light of PI's incentives.

The Waypoint Game

There are two players, the principal investigator (PI) and the Program Officer (PO). The focus of the game is on consideration of a research project X described previously. A good project yields a

desirable transformational outcome. A bad project has no chance of producing a transformational outcome. R's type is chosen by Nature and is not revealed to either player, however both players know the magnitude of p(G).

The goal of the PI is to have X funded.[3] PI receives a payoff of 1 if the PO funds X and 0 otherwise, less costs incurred by pursuing waypoints. PO's goal is to maximize the probability of achieving transformational research results, either by funding project X or an alternative, Project A. Specifically, PO receives a payoff of 1 if their chosen project is ultimately transformational and 0 otherwise. The likelihood of project A yielding transformational results is (a), and the probability that it does not is (1-a).

The game proceeds as follows:

Step 1: PI decides to pursue a waypoint, a decision that is private information. Waypoint pursuit forces PI to incur a cost c ($c < 1$).[4] However, pursuit yields one of two signals about R, H or ~H. The probability that either signal is observed depends on g as well as the probability that each type (G, ~G) yields each signal (H, ~H). Assuming that PI pursues the waypoint, the probability of observing different signals is set as follows:

$$p(H|G) = q \qquad p(\sim H|G) = (1 - q)$$

$$p(\sim H|\sim G) = v \qquad p(H|\sim G) = (1 - v)$$

where p and v are between .5 and 1, which allows us to calculate the probability of seeing each potential signal:

$$p(H) = p(H|G)*p(G)+p(H|\sim G)*p(\sim G) = (q)*(g)+(1-v)*(1-g)$$

$$p(\sim H) = 1 - p(H)$$

[3] Future work will explore the implications of PI having multiple motivations, such as an interest in funding and an interest in innovation

[4] This specification reflects the assumption that resources devoted to pursuing waypoints (publications, patents, etc.) represent a tax on the research process. One possible elaboration is to assume (either as a substitute or in addition to the costs specified here that pursuing waypoints can reduce the chances of generating transformational research outcomes.

and allows us to calculate how PO will update their beliefs about X given these signals:

$$p(G|H) = [p(H|G)*p(G)]/[p(H|G)*p(G) + p(H|\sim G)*p(\sim G)]$$

$$= [(q)*(g)]/[(q)*(g) + (1-v)*(1-g)]$$

$$p(G|\sim H) = [p(\sim H|G)*p(G)]/[p(\sim H|G)*p(G)+p(\sim H|\sim G)*p(\sim G)]$$

$$= (1–q)*(g)/[(1–q)*(g)+(v)*(1-\square g)]$$

Step 2: PO decides whether to fund X or fund A.

Let p(G*) be PO's estimate of the probability that project X is good, given PI's waypoint decision and the resulting signal. If H is observed, p(G*) = p(G|H). If ~H is observed, PO's belief is either that p(G*) = p(G|~H) if PI is thought to have pursued the waypoint, or p(G*) = (g) if PI is thought to have not pursued. PO's goal is to maximize the likelihood of transformational results, and so funds R if $p(G^*) \geq p(A)$ and funds A otherwise.

Equilibrium Considerations

Given the nature of the waypoint game, the obvious method for solving the game is a perfect Bayesian equilibrium, where each player chooses optimally given expectations about the other player's choice and the information available to them at the point of decision. Put another way, PI will invest in pursuing a waypoint only if doing so is necessary to influence PO's choice – and if the changes of achieving the waypoint are high enough to outweigh the costs involved with doing so. Thus, the critical parameters are p(G), p(A), and c, along with p(G|H).

For example, when p(G) exceeds, p(A), PI has no incentive to pursue a waypoint – project G is promising enough that funding will be continued even if new information is received (moreover, it is possible that pursuit if W will fail, such that p(G|~H) < p(A), and PO will respond to the negative signal by withdrawing funding from G). Similarly, if p(G) < p(A) and p(G|H) < p(A), project G is doomed, as PO will switch from G to A even if H is achieved. Under these conditions PI again has no reason to pursue W.

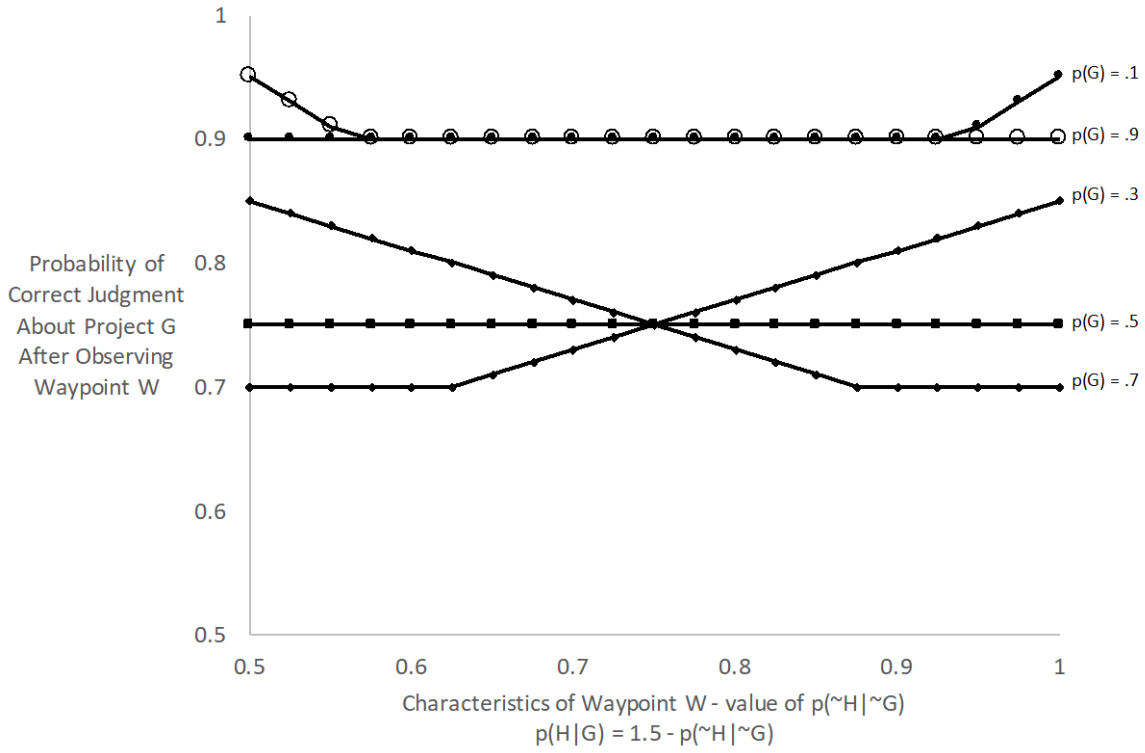
The potential for PI to pursue W is when (a) $p(G) < p(A)$, (b) $p(G|H) > p(A)$, and (c) $p(H) > c$. These three conditions imply that (a) without new information, PO will switch from G to A, (b), successfully achieving H will boost $p(G^*)$ such that PO will stay with G, and (c) the likelihood that pursuit of the waypoint is successful (and the benefits that accrue from success) is large enough to outweigh the costs of pursuit.

Equilibrium Behavior: Inferences and Expected Payoffs

The most interesting feature of the Waypoint game concerns the variation in PO's information and payoffs across different waypoints. Suppose, for example, that PO can choose any waypoint along the dotted line in figure 1, ranging from W_2, which has a zero chance of sending a false negative signal and a 50/50 chance of sending a false positive (from ~H), to W_3, which never sends a false positive but has a 50/50 chance of sending a false negative (from W), to all of the other waypoints along the dotted line, each with different values of $p(H|G)$ and $p(\sim H|\sim G)$.

Figure 3 shows that the quality of PO's inferences post-waypoint vary depending on which of these waypoints are used to assess G and, moreover, that the optimal waypoint for PO depends on the value of p(G). The horizontal axis of figure 4 gives the values of $p(\sim H|\sim G)$ and $p(H|G)$ for each potential waypoint along the dotted line.

Figure 3.
Quality of Information from Waypoints



The figure measures the accuracy of PO's post-waypoint judgment about G as a function of p(G) and the characteristics of the waypoint used to make the judgment. For example, suppose p(G) = .3 and PO uses a waypoint where p(H|G) = .875 and p(~H|~G) = .625 (this is the point in the p(G) =.3 plot where the line turns from horizontal to moving upward). Using this waypoint, PO will update p(G*) to .50 if H is observed and .08 if ~H is observed – put another way, given ~H, PO's best guess is that G is bad, and the likelihood that this guess is correct is .92; when H is observed, PO's best guess has a .5 chance of being right regardless of which one she chooses. After accounting for the likelihood of observing H and ~H, the odds of PO's judgment being correct is .70, which corresponds to the y-axis position of this point on the line.

Figure 3 shows that some waypoints are better than others in terms of informing PO. In the case discussed here, where p(G) = .3, if PO has a choice of all the waypoints along the dotted line, her best choice is the waypoint where p(~H|~G) = 1 and p(H|G) = .5; that is, a waypoint that minimizes false positives and maximizes false negatives. This result holds as long as p(G) < .50. In contrast, when p(G) > .50, PO maximizes her chances of making the right judgment about G by using the waypoint where p(H|G) = 1 and p(~H|~G) = .5, that is, a waypoint that maximizes false positives and minimizes false negatives.

At this point we are only at the beginning of analyzing equilibrium behavior. Initial analysis shows that PO's payoffs and choice hinge on the relative size of p(G) and p(A), and on the beliefs formed after observing w and ~H. Suppose PI pursues W, allowing PO to update their beliefs about G, p(G*). Given the current game structure, PO will choose G if p(G*) $\geq$ p(A) and A otherwise. Given this behavior, PO's expected payoff is p(G*) if they stay with G and p(A) if they switch. Since PO makes the last choice in this game, they will choose whichever option gives them the highest expected payoff, based on the value of p(G*) calculated after observing W. [5]

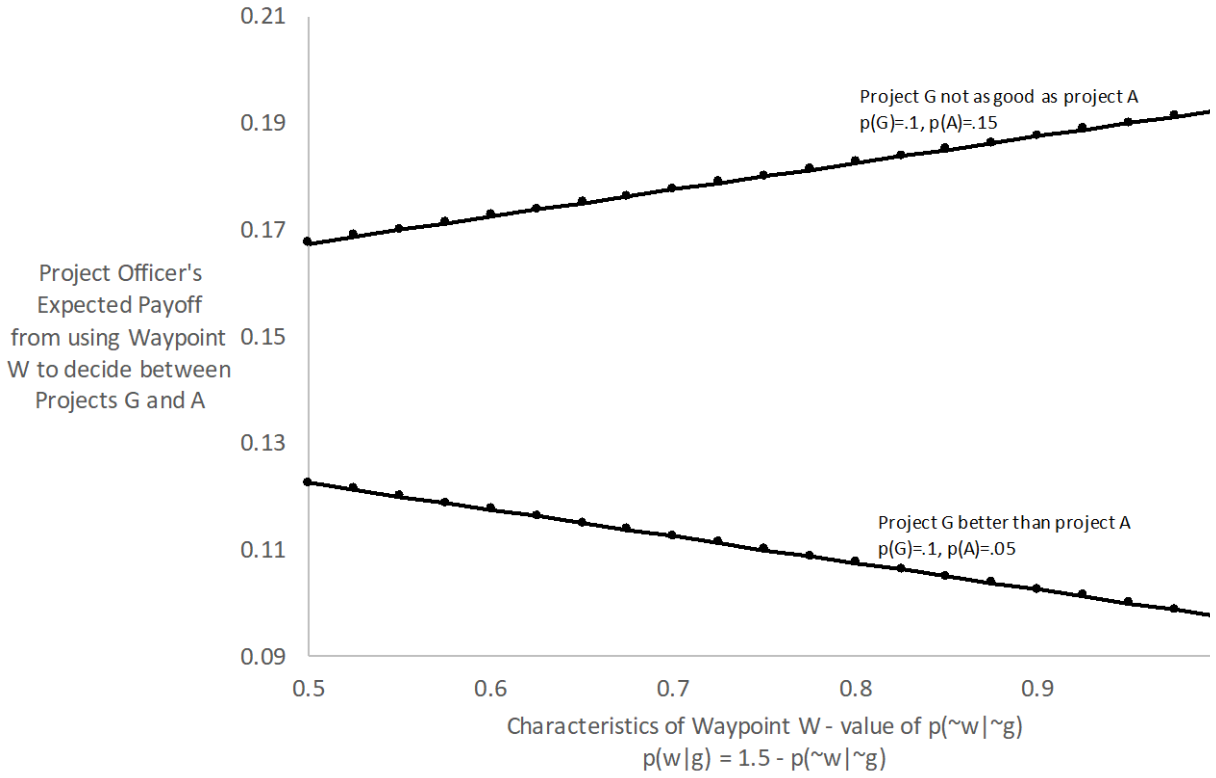
Figure 5 shows expected payoffs across the range of possible waypoints for two scenarios, one where p(G) = 1.5*p(A), and another where p(G) = .5*p(A).[6] The interesting result from figure 5 is that the choice of an optimal waypoint is always one of the extreme options, either the necessary or the sufficient waypoint. However, the question of which one is optimal (gives PO the highest expected payoff) depends on the relative sizes of p(G) and p(A): when p(G) exceeds p(A), the best option is a necessary waypoint; when the sign is reversed, the optimal choice is a sufficient waypoint.

---

[5] Suppose PI pursues W, allowing PO to update their beliefs about G, p(G*). Given the current game structure, PO will choose G if p(G*) $\geq$ p(A) and A otherwise. Given this behavior, PO's expected payoff is p(G*) if they stay with G and p(A) if they switch. Since PO makes the last choice in this game, they will choose whichever option gives them the highest expected payoff, based on the value of p(G*) calculated after observing W.

[6] We have looked at other payoff configurations, but thus far, changes in the magnitude of p(G) and p(A) shift the expected payoff line up and down, but the slope of the line depends on the relative side of the two probabilities as in the example.

Figure 4.
Choosing the Right Waypoint: Expected Payoffs



Implications: Choosing Benchmarks

The analysis provides guidance for decision-makers who must decide among possible waypoints or milestones for evaluating a research project. Our work shows that the optimal choice (in terms of maximizing the innovation rate) depends on the likelihood of innovations from the project under scrutiny and the alternate project that will receive funding if the default is canceled. When the current fundee has the higher (prior) probability, sufficient waypoints maximize the innovation rate; when the probability is lower, necessary waypoints are preferable. This result will need to be changed if we account for the value of innovations from each project – but the essence of the result is likely to remain in place. Under the circumstances captured by the Waypoint Game, the best

waypoint is to be found at the extremes of the feasible set, not in the middle. That is, the best milestone is either one that minimizes false positives, or minimizes false negatives.

Implications: Timing and Choice of Waypoints

The character of a waypoint may also chance over time. For example, we have described publications as akin to a necessary waypoint – a signal with a high false positive rate and a low false negative rate. But this description implicitly assume that sufficient time has elapsed to develop results. Consider, however, a very early-stage publication or similar finding from a drug development program. Late in the research cycle, this waypoint would be necessary. However, if it is achieved very early in the cycle, the inferences drawn from it could take the form of as sufficient waypoint – low false positive, high false negative. The implication for evaluators is that waypoints have different implications for project prospects at different points in the research cycle.

Implications: Political Control

For theories of political control, these results have a simple prediction: a program's ability to resist political pressures hinges on the availability of waypoints that meaningfully distinguish potential successes from potential failures. We expect that the set of potential waypoints, as well as their predictive power, will vary across different basic research programs. If so, programs with many highly-diagnostic waypoints available to evaluate projects will be better-able to resist congressional interventions compared to programs where there are fewer, less-diagnostic waypoints. For example, new or experimental programs, where the research process is poorly-understood, are ripe for political intervention, while established programs, where research is more likely to follow a predictable path, will be less vulnerable.