The Empirical Framework for the Evaluation of Research Funding

Julia Lane, New York University

Nicolas Jeanrenaud, New York University
Contents
1. Introduction ............................................................................................................................. 3
2. Overview of empirical approach ............................................................................................. 3
3. Empirical approaches .............................................................................................................. 5
4. Measurement .......................................................................................................................... 8
5. Selection Bias and Counterfactuals ....................................................................................... 10
6. Summary ................................................................................................................................ 14
7. References ............................................................................................................................. 16
8. Appendix on Propensity Score Matching ............................................................................... 18
1. Introduction

The empirical evaluation of science is, in principle, no different from the evaluation of any other kind of intervention. Any empirical evaluation must be guided by a theory of change that spells out how the funding intervention is expected to change activities, outputs and outcomes. Measurement issues are critical. Selection bias must be addressed - since research investments are typically based on a set of selection criteria, the approach must adjust for non-random participation. And the appropriate comparison group (or groups) must be carefully identified.

This white paper describes the empirical challenges and provides an overview of how they have been addressed, both within the field of research funding and in other domain areas.

2. Overview of empirical approach

The standard impact evaluation framework is to determine what is the impact (\( \Delta \)) or causal effect of a program (\( P \)) on an outcome of interest (\( Y \))

\[
\Delta = (Y|P=1) - (Y|P=0)
\]

In other words, the causal impact (\( \Delta \)) of a program (\( P \)) on an outcome (\( Y \)) is the difference between the outcome (\( Y \)) with the program (\( Y=1 \)) and the same outcome (\( Y \)) without the program (when \( P=0 \)). The approach is to describe outputs, confounding factors, and develop a counterfactual. The Government Accountability Office (2012) defines impact evaluation as follows:

*Impact evaluation* is a form of outcome evaluation that assesses the net effect of a program (or its true effectiveness) by comparing the observed outcomes to an estimate of what would have happened in the absence of the program. While outcome measures can be incorporated into ongoing performance monitoring systems, evaluation studies are usually required to assess program net impacts. P. 16. (2)
The most important empirical step to write down how to measure a theory of change (3). This is a logical chain that outlines a sequence of inputs, activities, and outputs for which a program is directly responsible and identifies pathways through which impacts are achieved (see Figure 1).

**Figure 1: Source Martinez et al. (1)**

In the case of research funding, one possible approach is to write down the hypothesized relationship (more precisely, two groups of interrelated regression equations) to illustrate the approach more formally\(^1\)

\[
Y_{it}^{(1)} = Y_{it}^{(2)}\alpha + X_{it}^{(1)}\lambda + \epsilon_{it} \quad \text{and} \quad Y_{it}^{(2)} = Z_{it}\beta + X_{it}^{(2)}\mu + \eta_{it},
\]

where the subscripts \(i\) and \(t\) denote research networks and time, respectively. \(Y^{(1)}\) is the output variable, or the creation, transmission and adoption of ideas. \(Y^{(2)}\) is the network structure. Both variables are determined by a set of control variables \(X^{(1)}\) and \(X^{(2)}\) that can overlap, such as scientific infrastructure or the prior state of knowledge. The variable of key interest, \(Z\), is funding. The parameters of interest include \(\alpha\), the efficiency with which research networks convert their activities into the transmission and adoption of ideas, and \(\beta\), the efficiency whereby

---

\(^{1}\) This specification draws on joint work of Lane with Jacques Mairesse and Paula Stephan
funding creates and sustains those networks. The remaining parameters, $\lambda$ and $\mu$, reflect the importance of the contribution of the control factors; while $\varepsilon$ and $\eta$ stand for unobserved factors, including pure serendipity, and errors of measurement and specification.

Any impact evaluation must be conducted by an external group to maintain objectivity and credibility. Of course, that process cannot be divorced from the operational rules and it is particularly important to assess the rules of program implementation or program fidelity. The evaluation team should work with the program implementers to develop the appropriate evaluation design and ensure that program implementation and evaluation are well coordinated so that one does not compromise the other. (1)

3. Empirical approaches

The empirical literature examining the impact of research funding is well surveyed by Stephan, (4, 5) and in a variety of reports (6–8). Table 1 provides a summary of early approaches, including surveys, case studies, bibliometrics, econometrics and statistical analyses, content analysis, and expert judgment, which have been used in program evaluation.
Previous research on the economic and scientific impact of scientific funding has used information on intermediary outputs, notably patents and publication, to “proxy” actual impact\(^9\), \(^{10}\). Yet another stream of research has explored the determinants of academic entrepreneurship, whereby economic impact is achieved by the participation of academic faculty in the founding of university technology based start-up firms \((11–13)\). Recent research has examined the channels
through which impact occurs including contract research, joint research between universities and industry and consulting (14–17). This work has made progress in accounting for the wide spectrum of activities that do not leave any intellectual property traces.

There has been substantial research on health outcomes as well. An international network hosted by the National Bureau of Economic Research and the Institute for Fiscal Studies (of which Lane is a member) has spent the past four years building an empirical framework.

As that work points out

“the link between the inputs and ultimate outputs of the research process is complex. Consider just a few examples on which there is already existing related research in the broad area of medical research:  Has research led to health improvements that are worth more than the costs of providing them? Are people aging healthier, and if so how have medical and social research affected healthy aging? In what ways are public and private research substitutable, and in what ways are they complementary? What metrics can be used to value the output of the research process? In all these areas there is more that can be known, the existing results and findings can be brought together, and there is also a need for sharing of information across fields and disciplines with regard to the theoretical and empirical methodologies that are currently applied. Other relevant questions have received less attention. For example, depending on the availability of data and suitably analytical models, further questions might relate to: how to treat uncertainty over research outcomes, the potential returns to failure and the difference between ex-ante and ex-post valuations of research; the effect of health research on the evolution of health over the entire life-cycle and how this affects our view of its value; the different conceptual models required to value (perhaps speculative) research fields as opposed to more specific and concrete technologies or treatments within a field; the macroeconomic gains to a healthy and thriving medical research sector; how one might build distributional considerations into an understanding of the social value of medical research”

(\url{http://www.nber.org/aging/valmed/research_ifs.html})

The same work provides a useful overview of current literature with a specific health focus.

While previous research has been valuable in informing the determinants of scientists’ economic and social activities such as the above, there has been a lesser focus on the production of human capital within publicly funded science. A crucial way in which public science expenditure adds value to economy and society particularly in the medium and long-terms is via

\[\text{http://www.nber.org/aging/valmed/}\]
the “production” of highly skilled people(18, 19). For instance, it is widely accepted that the
development of regions is bound up with the quality of the human capital that resides in them, or
is attracted to them(20), and does not depend on the knowledge distributed in scientific papers.

The modern empirical approach is to map the structure of scientific collaborations and
the link to new ideas and new human capital by using “big data” techniques and build new data
at the project level(21, 22).

4. Measurement

Measurement issues are always cited in any impact evaluation; the same is true in the
context of science policy3. In particular, there are issues of attribution: A given innovation may
draw upon multiple research projects and many other inputs before market and social effects are
realized. There are challenges in identifying who the beneficiaries of research are, so
appropriability is an issue. Finally, as with much human activity, particularly with risky
activities, the distribution of impacts is likely to be skewed, with only a few projects having big
effects.

Probably the most challenging issue is that of timing, since the causal chain is long and
complex. High commercial value is often generated over decades, with inputs and collaborations
with private industry, and cross fertilization from multiple areas of research. It is important to
determine whether sufficient time has passed to actually observe outcomes. If analysis is
premature, a program might be deemed ineffective, even though it is likely to yield significant
results in the future, and it might be critical for another program’s ongoing success. These types
of externalities are extremely difficult to measure, and oftentimes require detailed case studies to
observe and appreciate the potentially transformational yet veiled impacts.

---

3 This section draws heavily on work by Kaye Husbands Fealing (37)
Historical tracing is another method to be considered: this relies on context-specific analysis, expert evaluation and it is time intensive. Figure 2 (below) shows the impacts of several areas of basic research, some of which originated at universities.

**Figure 2** Examples of the contributions of federally supported fundamental research to the creation of IT sectors, firms, and products with large economic impacts.(7)

Regardless of the approach used, it is particularly important to analyze the processes established to achieve the particular goals of the program implementation. There are three different levels for this analysis. The first is what is called program fidelity: does the program coherently follow the established general goals; if this coherence does not exist, it is very difficult, if not impossible to achieve the desired results. The second is whether there is a rational link between the procedures and outcomes as well: whether there is a coherent theory that links
goals with results. The third is understanding the relative importance of the different inputs and their relationship to the processes and outcomes.

5. Selection Bias and Counterfactuals
The identification of a counterfactual is critical to identifying the effect of an intervention ($\Delta$),

$$\Delta = (Y \mid P=1) - (Y\mid P=0)$$

as Figure 3 makes clear. If the chosen counterfactual is B, the estimated impact of an agricultural intervention is 100; if C, the effect is half that, if D, 50% more.

![Figure 3.3 Before-and-After Estimates of a Microfinance Program](image)

**Figure 3** The effect of different counterfactuals on the estimation of impact

It is important to identify the nature of the intervention and the target unit of analysis, which ties back into correctly specifying the theory of change. In particular, is the intervention intended to be an individual grant, an entire set of programmatic interventions, an entire research portfolio or research funding in general? Similarly, is the target unit of analysis intended to be one individual researcher, an institution, a narrow research field or a broad research field?
Research funding is typically predicated on a peer review process that funds the “best” research, which creates a fundamental evaluation problem due to selection bias (23). One way to adjust for selection bias is to institute randomized controlled trials, which are an important way in which program evaluations are implemented in other fields. Azoulay has argued strongly for randomized controlled trials (24).

Another way to address the problem is to identify natural experiments. These are also rare in research, although some work is now being done using either clever instruments or true exogenous shocks (25, 26), where “before” and “after” outcomes are compared.

Another approach would be to use propensity score matching of individuals who received research funding and compare them to individuals who didn’t. This approach has been extensively applied since its introduction by Rosenbaum and Rubin (27, 28), and essentially groups similar individuals based on the propensity to be treated (see Smith and Caliendo & Kopeinig, (29, 30) for good discussions). However, in the case of comparing research portfolios across countries, propensity score matching is difficult to implement.

A related important methodological approach is “differences in differences”. This methodology is particularly useful when it is not possible to observe directly the characteristic of a population (persons, firms, etc) participating in a program. The starting point is to have a concrete population of agents receiving funding to achieve a particular objective (to increase educational results, economic competitiveness and so on). The first step is to find objective information about the performance of the subjects (for example, publications, patents, economic

---


5 A technical summary is provided in the appendix.
value..) along a sufficient number of years before and after the period in which they have received the aid. There a number of changes can be observed. The methodological problem is how to isolate the effect of the program from other confounding factors which might affect selection into the problem but which are not directly observable. The standard approach is to construct a statistical equivalent comparison group with the same structural characteristics as the actual sample(31, 32). The next task is to capture information about their performance for the same period of time. The assumption is that the behavior of the control sample would predict the behavior of the actual sample in absence of the funding; so, the differences between the control group performance and the actual sample is inferred to be due to the funding intervention.

A closely related approach to both of these is to create synthetic cohorts – this has been used to compare different interventions – such as California’s tobacco control program on state-wide smoking rates relative to other states (33) and the effect of Basque terrorism in Spain relative to other regions (34). This approach has been called arguably the most important innovation in the policy evaluation literature in the last 15 years(35). Simply put, rather than choosing one entity to be the comparison group, the synthetic control approach uses a weighted average of the set of controls.

In the United States, econometric analysis has been used to measure impacts, given appropriate discontinuity designs. The list of illustrative studies is quite long in this arena. A recent study shows the impact of students who immigrate to the U.S. on innovation. Teasing out these impacts is skillfully done using instrumental variables procedure(14). Econometric analysis is used to assess the impact of National Science Foundation grants on research output of individual researchers primarily at universities. They use data on (quality-adjusted) publications for principal investigators five years prior to getting an NSF grant and five year after the PIs
received the NSF grant. While grants had no impact on the productivity of senior researchers, it
did affect junior researchers: the regression discontinuity design used by Arora and
Gambardella—along with the focus of analysis on the grant or the PI—allows for the
counterfactual question to be directly addressed(15).

Before-and-after types of econometric analysis have also been used to determine the
effect of changes in economic policy. One study examined whether a change in policy by the
National Institutes of Health to allow greater sharing of intellectual property between researchers
affected the amount of experimentation done on basic science and applications leading to
commercialization of products and processes. They find that easing limitations on intellectual
property rights allowed for greater diversity in experimentation(16) Another observed the
“special collections” of biological materials at biological resource centers, to determine whether
those institutions played a role in enhancing knowledge creation. Again, there is a clear period
in time when a change was made and the impact of the change was observed at the micro
level(17). The implication for research organizations here is that tracing specific research
activities through to outputs and observing natural experiments is one means of isolating impacts
and understanding causal effects. These two cases also show that research organizations can
develop policies that enhance knowledge generation, knowledge flows and knowledge transfer
outside of the institution, and that those policies can have real, measurable impacts on socio-
economic outcomes in addition to the generation of new knowledge.

While the answer will depend on the interest of the research funder, one possible
approach has been developed in joint work with Kaye Husbands Fealing and Matt Ross(36). This
study investigated the impact of food safety research, which is funded by multiple agencies –
National Science Fundation (NSF), National Institutes of Health (NIH) and US Department of
Agriculture (USDA). In order to examine the impact of the research, it is critical to identify the comparison group – should it be investments in computer science and engineering in the case of NSF programs, investments in genomics research in the case of NIH programs, or investments in nutrition in the case of USDA? Clearly the choice of baseline affects the estimated impact. In this case, the theory of change was that investments in food safety research were intended to entice researchers trained in closely comparable fields to think about food safety problems. Hence, comparing outcomes of computer scientists and engineers would be inappropriate.

Operationally, text analysis was used to identify both food safety grants and grants that were “not food safety” as well as grants associated with each funding program – the NSF, NIH, and USDA. Then “food safety funding programs” were characterized in two ways – those that have funded at least one award that was classified food safety (called “food safety funding program”) and those that had at least 5 percent of their funding portfolio in food safety (called “intense food safety funding program”). The researchers then identified three possible comparison groups. One comprised grants that were funded by USDA that were “not food safety.” The second consists of grants that were “not food safety” in a food safety funding program. The third comprises grants that are “not food safety” in an “intense food safety funding program.” In this way, relatively similar research portfolios were compared to each other. A different theory of change would have led to a different approach.

6. Summary

There is a substantial evaluation literature that can be used to inform the empirical approach. A number of key issues must be addressed in any empirical approach. The first is conceptual: the evaluation must be guided by a theory of change - the reason for the intervention – in order to measure the effects on the expected outcomes. The second is to address
measurement issues. And the third is operational: since any evaluation must compare the intervention to a comparable counterfactual, and funding agencies select participants based on a set of selection criteria, the approach must carefully consider selection bias, adjust for non-random participation, and identify appropriate comparison groups.
7. References

19. N. Zolas et al., Wrapping it up in a person: Tracing flows from funded research into the
32. J. Smith, A. Sweetman, in *Strengthening Evidence Based Policy in the Australian Federation, Volume 1: Proceedings*, P. Commission, Ed. (Canberra, Australia, 2010).
8. Appendix on Propensity Score Matching\textsuperscript{6}

Consider the model:

\[ y = \alpha + \gamma D + \epsilon \]

where \( D \) is an indicator variable for research funding and \( y \) is the outcome. The parameter \( \gamma \) is the coefficient of interest. The literature considers three parameters of interest: the average of treatment effect or \( \Delta_{ATE} \), the average impact of treatment on the treated or \( \Delta_{TT} \), and the average impact of treatment on the nontreated or \( \Delta_{TN} \). We let \( p(x^i) \) be the propensity score at \( x = x^i \), where \( x \) is our vector of covariates. The weights that allow \( \gamma \) to estimate the various \( \Delta \)'s are given in Table A1 under the assumption that we have a probability sample of the relevant population. For this project, we are concerned with the performance of those funded by research so we wish to focus the parameter \( \Delta_{TT} \). In this case, we weight observations in the comparison group with the weight \( \frac{p(x^i)}{\sqrt{1 - p(x^i)}} \). The role of these weights in the regression is to ensure that the distribution of covariates in the comparison group is identical to the distribution of covariates in the treatment group. This means, of course, that our covariates are statistically independent to our indicator for research funding, \( D \).

Now, when one has covariates that are statistically independent of the research funding indicator, they do not need to be entered into the regression, since they will have no effect on

\textsuperscript{6} Drawn from Black and Lane(38)
the estimate of $\gamma$. (They might be entered into the regression to reduce residual variance, but
that will be discussed later.)

If we let the covariates be discrete, we can make the intuition a bit easier. Thus, assume
the data are discrete so that the data forms cells and there will be a $p_j$ associated with each
cell that tells us the fraction of individuals within the cells that are funded by research. The
least squares objective function is:

$$L = \sum_{i=1}^{n} \left( \frac{1}{N_{nit,w}} w_i (y - \alpha - \gamma D) \right)^2$$

where $w_i$ is the weight and $N_{nit,w}$ is the sum of the weights. Some algebra shows that

$$\gamma_{w}^{OLS} = \bar{Y}_{1,w} - \bar{Y}_{0,w}$$

where $\bar{Y}_{D,w}$ is the weighted mean conditional on whether the respondent was in the military or
not. This may be rewritten as:

$$\gamma_{w}^{OLS} = \frac{1}{N_{nit,w}} \sum_{p_j} \sum_{x=p(x)-p_j} \left( w_{j,D=1} D Y_{i,D=1} - w_{j,D=0} (1-D) Y_{0,i} \right)$$

where $i$ indexes the individual, or

$$\gamma_{w}^{OLS} = \frac{1}{N_{nit,w}} \sum_{p_j} \left( w_{j,D=1} n_{j,D=1} \bar{Y}_{1,j} - w_{j,D=0} n_{j,D=0} \bar{Y}_{0,j} \right)$$

For $\Lambda_{IT}$, the weights are $w_{j,D=1} = 1$ and $w_{j,D=0} = \frac{n_{j,D=1}}{n_{j,D=0}}$ and

$$N_{nit,w} = \sum_{p_j} n_{j,D=1}$$

so we may write:

$$\gamma_{w}^{OLS} = \sum_{p_j} \left( \frac{1}{n_{j,D=1}} \sum_{p_j} \left( n_{j,D=1} \bar{Y}_{1,j} - \bar{Y}_{0,j} \right) \right)$$
where the subscript \( j \) indexes the data cell. For a particular cell, we simply compare
\[
\left( \bar{Y}_{1,j} - \bar{Y}_{0,j} \right)
\]
the outcomes of those funded by research with the comparison group. To obtain
the average impact of treatment, we then weight each cell by its relative size among the
research funded sample \( \sum_{\forall p_j} n_{j,D=1} \). Because the estimator compares only people with identical
values of the covariates, the covariates have no impact on the estimation.

Of course, with continuous data, we cannot formally express the estimator as a simple
difference of means but the intuition still holds: we have made the covariates independent of
the indicator for research funding so we are assured they do not confound our estimation.

Finally, because we are interested in reducing the residual variation to improve
the power of our statistical tests, we do not estimate the model
\[
y = \alpha + \gamma D + \varepsilon
\]
but estimate the more standard looking regression model
\[
y = x\beta + \gamma D + \varepsilon .
\]
Because the covariates \( x \) are independent of the research funding indicator \( D \) by
construction, we do not have to worry whether we have specified the regression model
correctly. All we wish to do by including our covariates is to reduce the variance of our
regression error and make our various hypotheses tests more powerful.

<table>
<thead>
<tr>
<th>Table A1: Weights for Nonparametric Estimates of Evaluation Parameters when Data are a Probability Sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parameter</td>
</tr>
<tr>
<td>---------------------</td>
</tr>
<tr>
<td>( \Delta_{dTE} )</td>
</tr>
</tbody>
</table>

HELIOS workshop September 3, 2018
INCa/ITMO Cancer-Aviesan
| $\Delta_{TT}$ | $\frac{1}{p(x')}$ | $\frac{p(x')}{[1-p(x')] \cdot 1}$ |
| $\Delta_{TN}$ | $\frac{1-p(x')}{p(x')}$ | |