BUILDING PROGRAM EVALUATION INTO THE DESIGN OF PUBLIC RESEARCH SUPPORT PROGRAMS

Adam B. Jaffe
Brandeis University and National Bureau of Economic Research

January 2002

Forthcoming, ***Oxford Review of Economic Policy***

An earlier version of this paper was presented at TECHNOLOGICAL POLICY AND INNOVATION: Economic and Historical Perspectives, Paris, November 2000. The research underlying this paper was supported by the U.S. National Science Foundation. Lori Snyder provided valuable research assistance. The views expressed are those of the author.

I. INTRODUCTION

It is widely accepted that, in the absence of policy intervention, the social rate of return to R&D expenditure exceeds the private rate, leading to a socially suboptimal rate of investment in R&D (Guellec and van Pottelsberghe, 2000). Indeed, empirical evidence suggests that, even given the public support of R&D typical in OECD countries, the social rate of investment in R&D remains suboptimal (Griliches, 1992; Jones and Williams, 1998). All of this suggests that finding ways to foster increased investment in R&D ought to be a significant public policy concern.

Most OECD countries subsidize R&D via tax treatment that makes the after-tax cost of R&D considerably lower than that of other forms of investment (Guellec and van Pottelsberghe, 2000; Hall and Van Reenen, 2000). In addition, most countries have publicly funded research grant programs that attempt to funnel public resources directly to R&D projects that are believed to have particularly large social benefits. Such research grant programs include those that support basic scientific research, R&D aimed at particular technical objectives of importance to the government (e.g., defense, health, environment), and 'pre-competitive' R&D intended to generate large spillovers, often with a collaborative component.

Despite the prevalence of such programs, however, there is little consensus about their effectiveness. Although there are a small number of studies that seem to demonstrate significant social returns to particular programs, there remain serious methodological questions about these findings (Klette, Møen and Griliches, 2000). In the U.S., in particular, there remains a significant and politically important suspicion about the desirability of public grants for the support of commercial R&D (Yager and Schmidt, 1997). At a conceptual level, there are two basic reasons why one might believe that public research grants would be socially ineffective, despite the existence of excess social returns to R&D. First, firms and other entities may not be as careful in their use of other people's money as they are of their own, and hence may waste it or use it unproductively. Second, public support may 'crowd out' private support, meaning that even if the public resources are used productively, there is no net increase in social investment in R&D. Given that public resources must be raised via socially costly revenue mechanisms, society is worse off if total R&D investment remains the same while public funding replaces private funding.

Much of the political debate surrounding such programs remains at the level of ideology. Opponents question at a conceptual level how government programs can pick 'winners and losers' without interfering in market processes to an undesirable extent, and point to expensive fiascoes such as the program to develop synthetic fossil fuels in the U.S. in the 1970s and 1980s. Supporters rely on the already mentioned theoretical case for inadequate private incentives, and point to important examples of socially valuable technology that were produced or midwifed by the government, such as commercial jetliners, communications satellites, and the internet.

I am not naïve about the political importance and staying power of ideology. Yet as social scientists we have an obligation to try to bring facts to bear on these debates. Given the evidence of excess social returns to R&D, combined with the questions about the effectiveness of public grants in increasing total R&D productively, the social productivity of these programs is fundamentally an empirical question. While we have made significant progress in providing some answers to this empirical question, I think we can do better.

To my knowledge, all of the empirical work evaluating the effectiveness of these programs has been what I will call 'after-the-fact' evaluation, by which I mean an evaluation in which a researcher comes along sometime after a set of grants has been made, and attempts to infer the effect of those grants using observational data collected at that time. In this paper, I explore the possibility of producing more compelling empirical evaluations by having the grant agencies anticipate the need for such evaluation and build certain features into the grant process to facilitate later evaluation. In the U.S., the granting agencies now have an incentive to engage in this sort of activity because the Government Performance and Results Act of 1993 (P.L. 103-62, 'GPRA') requires all agencies to report systematically to the Office of Management and Budget on the 'outputs' and 'outcomes' of their programs. After initially resisting the applicability of this law to their activities, research agencies are increasingly trying to figure out how to satisfy its mandates (Cozzens, 1999).

The focus of this paper is on how such program design for evaluation could produce data that would allow researchers to deal more effectively with the selectivity bias that is likely to plague any after-the-fact efforts to measure the effect of a research grant program. I will discuss the possibility both of experimental designs, in which some grant decisions are made randomly, and of other, possibly less intrusive ways to structure grant decisions that might mitigate selectivity bias in subsequent evaluations.

To be sure, selectivity bias is not the only problem that has to be solved in order to undertake convincing empirical evaluation of research grant programs. Other important problems—which this paper will not address—include: (1) how to measure research output of supported research entities; (2) how to measure the spillover benefits of funded research enjoyed by entities other than those that are supported; and (3) how to measure 'transformational' impacts whereby public support changes the nature of the research infrastructure, with possibly important long-lasting effects.[1] More generally all of the discussion in this paper focuses on the marginal, short-run or 'partial equilibrium' effect of a program. These modes of analysis cannot capture the long-run, general equilibrium impacts that an overall program has on the economy. This means that any estimates of program effectiveness that come out of these analyses would be useful for making decisions on program modification, expansion, or contraction at the margin, but would not be relevant in evaluating large changes or eliminating entire programs

The organization of the paper is as follows. The next section lays out the selectivity bias problem, and summarizes what approaches are available to deal with it in the context of 'after-the-fact' evaluation of research grant programs. The third section discusses how grant programs might be changed to facilitate subsequent evaluation. I discuss the use of 'randomization' to generate 'experimental' data, as well as an alternative approach, based on what statisticians call the 'regression discontinuity' design. This approach would require agencies to make only minor changes in their selection process, but would still offer significant benefits in allowing later evaluators to control for selection bias. The fourth section briefly discusses the so-called 'additionality' question: whether or not public research support increases grantees' total research spending. The final section provides concluding comments, including brief consideration of the political economy of better program evaluation.

II. THE SELECTION BIAS PROBLEM

A. A Canonical Research Grant Program

To begin, it is useful to have in mind a concrete, if somewhat stylized picture of how the research grant process works. Consider a public agency that disburses money for research on the following basis:

1. A legislative or higher-level executive agency establishes a budget of money available to be spent on a particular kind of research or research in a given substantive area.

[1] These and other broader issues in research program evaluation are discussed in Jaffe, 1998; Popper, 1999; Klette, Møen and Griliches, 2000, and Georghiou and Roessner, 2000.

2. Non-governmental entities apply for this money by writing and submitting substantive research proposals.[2]

3. The agency solicits from known outside experts in the field reviews or reports evaluating the proposals, possibly including a subjective numerical score for each proposal.

4. A committee or 'panel' organized by the agency meets, reviews the proposals and the external referee reports, and ranks the proposals in terms of priority for funding. This ranking may be incomplete, i.e., proposals may be classified into groups assigned different priorities, but within which no ranking is made.

5. An agency official then decides which proposals to fund, and how much money to grant to each applicant. This decision is based on the recommendations of the panel, the official's own judgement about the proposals, and other criteria not related to proposal quality, such as diversity of gender, race, geography and type of institution, or balancing of the grant portfolio by scientific fields.

6. Successful applicants are funded in the form of a 'grant', the distinguishing feature of which is that receipt of the funds is not conditional on the production of specified research outputs.

B. The Selection Bias Problem

The selection problem that arises in attempting to assess the impact of this kind of program is widely recognized. Klette, Møen and Griliches (2000) provide a useful overview, and Heckman, *et al.* (1998) give a rigorous statistical treatment. As a basis for discussion, consider the following version of the standard model:

$$Y_{it} = \beta_i D_i + \lambda X_{it} + \alpha_i + \mu_t + \omega_{it} + \varepsilon_{it} \qquad (1)$$

where Y_{it} is the research output of applicant i in period t,[**3**] D_i is a dummy variable that is unity if individual i receives a grant, and β_i is the effect

[2] The submitting entities would be firms, in the case of commercial research support programs, or academic researchers or other non-profit researchers in the case of basic science support programs. While much of the literature has focused on support of commercial research, evaluating programs that support basic research raises many of the same conceptual issues. Some programs target narrower kinds of institutions, such as joint ventures or collaborative research groups. Again, many of the same issues arise in evaluation.

for applicant i of receiving a grant. X_{it} is a vector of observable determinants of output (e.g., firm size, age of researcher). The unobservable determinants of research output are reflected by the last four terms. There is a time-invariant 'applicant effect' (α_i), and a time-period effect common to all applicants (μ_t). The usual error term that is assumed to be uncorrelated with the X's and D is represented by ε_{it}. The only non-standard entry is ω_{it} , which represents period- and applicant-specific variation in research productivity that is unobservable by the econometrician, but observable by the granting agency. It could represent agency officials' personal knowledge of the applicants. It could also represent the quality of the specific research project proposed to the agency, which the agency or other reviewers may be able to determine is better (or worse) than the time-invariant quality of the applicant itself captured by α_i. Note that while I use the word 'applicant' throughout to describe potential researchers, I assume that the equation applies to a population of potential applicants that includes some that never actually apply for funds.

Because the effect of the grant program is allowed to vary by applicant, our goal is to measure the average impact. Further, this average will be different for different groups of actual or potential applicants. For the purpose of benefit/cost analysis, we would like to know $E(\beta_i | D)$, the average effect of the grant program for those entities receiving grants. Note that, at this point, the impact of the program is associated with a dichotomous grant/no grant condition; I return below to the question of whether the magnitude of the research grant matters. In this way, the question being examined is the classic one of determining the effectiveness of a 'treatment' that is given to a non-random fraction of some population. We wish to determine the average effect of treatment on the treated group.

The obvious way to do this is to estimate some version of Equation (1) on a sample of applicants who did and did not receive grant funding, and use the regression coefficient on the treatment dummy as our measure of the treatment effect. I now consider what kinds of regression analyses of this general form might yield 'good' estimates of the average treatment effect.

I presume that the agency chooses whom to fund by attempting to maximize the impact of its funding, subject to its budget constraint and conditional on the information it possesses. If the world is described by

[3] I note again that I am ignoring the problem of how one actually measures research output. A related issue that I also ignore is the likelihood of long and variable lags between the receipt of the research grant and its effect, if any, on research output.

Equation (1), this problem reduces to ordering the projects according to their β_i's and choosing as many of the highest β_i-projects as fit in the budget. Note that, in and of itself, such selection does not create any selectivity bias. If the β_i are uncorrelated with everything else, we could estimate Equation (1) by Ordinary Least Squares, and the coefficient on the treatment dummy would tell us $E(\beta_i | D)$, which is what we want.[4]

The selection bias problem arises because we presume that β_i is correlated with α_i and ω_{it} across i. That is, the projects that are the best candidates for funding—in the sense of maximizing the impact of public support—are also the projects that would have the largest expected output in the absence of funding.[5] This means that selection on β_i makes $E(\alpha_i + \omega_{it} | D) > 0$, which biases the regression estimate of $E(\beta_i | D)$.[6]

C. Empirical Strategies for Mitigating Selection Bias in 'After-the-Fact' Evaluations

Regression with controls. The simplest approach to eliminating correlation between D_i and the error term is to include in the regression variables believed to 'control for' the unobserved effects. For example, Arora and Gambardella (1998) estimate a version of Equation (1) for all economists who applied for NSF funding over a five-year period, some of whom were funded and some of whom were not. They used impact-weighted publications in a five-year window following the grant decision as a measure of research output. They included in the regression as a control for α_i the impact-weighted publications in the five years prior to the grant decision. They also included, as a control for ω_{it}, the average outside-reviewer score received by the proposals.

There is, of course, no way to know whether any such set of controls adequately represents the information possessed by the funding agency. This is particularly true with relation to the quality of the different

[4] Of course, we could not determine the unconditional or population average for the treatment effect, but that average has little policy relevance because we don't contemplate 'treating' the entire population.

[5] The extent of *ex-ante* correlation between the treatment effect and the productivity in the absence of treatment may vary depending on the agency mission. But it is hard to imagine a situation in which one could be reasonably confident that the treatment effect is independent of baseline productivity. Note that the treatment effect could be *negatively* correlated with expected performance in the absence of treatment, if some applicants 'need' public support because their performance would suffer badly without it. In this case, selectivity bias would lead the regression to underestimate the average treatment impact.

[6] For a formal and detailed discussion, see Heckman, *et al.* (1998). Kauko (1996) discusses the problem of selection bias in evaluation of R&D subsidy programs.

proposals, as distinct from the quality of the applicant. Indeed, if it were possible to predict agency funding decisions on the basis of a regression on observable characteristics, one would have to question the rather significant social resources that these agencies typically expend on their subjective decision processes.

Matched samples of treated and untreated entities. An approach that is closely related to regression with controls is to compare treated entities to a sample of untreated entities that is drawn to resemble, as closely as possible, the treated entities, with respect to all observable characteristics that are believed to be correlated with likely performance (Brown, Curlee and Elliott, 1995).[7] Again, the difficulty is that it seems unlikely that similarity with respect to these observable attributes is sufficient to avoid the likelihood that the expected performance of the selected group would exceed that of the control group even without treatment.

Fixed effects or 'difference in differences'. The time-invariant unobservable α_i can be eliminated from Equation (1) by taking the difference in performance after treatment as compared to the performance before treatment. If such a difference is taken also for untreated entities, then common time effects μ_t are also eliminated by using the difference between the average before/after difference for the two groups (conditional on the difference in the X's) as the estimate of the treatment effect. Compared to the above approaches, this approach has the advantage of eliminating any need to find observable correlates of the unobserved productivity difference. In a recent careful study, for example, Branstetter and Sakakibara (2000) show that Japanese funding of research consortia increased the research output of the participating firms, in the sense that the output of these firms increased during and after their funding more than output of nonfunded firms increased over the same period.

The limitation of this approach is that it only controls for time-invariant unobservables. To the extent that the agency can and does evaluate the proposed *project* distinctly from the proposing *entity*, the resulting selection bias is not eliminated by differencing. In addition, one could imagine other sources of unobserved performance differences that vary across individuals and time. For example, applicants may decide to enter the grant competition when they have been enjoying unusually good (or bad?) recent performance. Any unobserved variation of this kind makes the differences estimator biased; differencing eliminates the

[7] Lerner (1999) used this approach to assess the impact of Small Business Innovation Research (SBIR) funding on the R&D of small firms.

time-invariant α_i, but introduces a new error related to the deviation in the previous period from the applicant's 'normal' performance. Indeed, depending on the relative magnitude of time-invariant and time-varying individual effects, differencing could produce estimates that are more biased than simple regression estimates.

Selection models/instrumental variables. The final approach relies on a model of the selection process to control explicitly for the conditional dependence of D_i on the unobservables. Identification of this approach derives from an exclusion restriction, which comes from either a variable that affects the probability of selection but does not affect performance, or an assumption about the functional form of the relationship between the unobservables and the dependent variable. Subject to the validity of this restriction, this approach provides valid estimates of the treatment effect regardless of the nature of the unobservables and the selection agency's knowledge of them.

The most familiar example of this approach is the latent variable model, in which it is assumed that selection occurs when an unobserved index surpasses some threshold value. The index or latent variable is assumed to depend on some observables plus an error drawn from a parametric distribution. If the determinants of the latent variable are all elements of the vector X in Equation (1), then the exclusion restriction that identifies the model is that the particular *function* of those variables created by their interplay with the parametric distribution of the error in the selection equation and the selection threshold is excluded from Equation (1).

The latent variable model is closely related to the instrumental variables approach, in which instruments that predict selection but not performance are used to estimate the effect of selection in Equation (1) consistently.[8] The difficulty in implementing this approach is in finding variables that affect selection that are not related to expected performance. The classic examples of this approach to the selection problem are ones in which institutional 'quirks' introduce observable correlates of selection that are not related to expected performance.[9] In the case of research funding programs, the most likely candidates for instruments are variations in the available budget and various kinds of 'affirmative action' in the selection process. Wallsten (2000) looked at a

[8] Conceptually, the latent variable approach can be thought of as using the parametric distribution for the latent error to create an instrument for selection even where there are no variables that predict selection but not performance.

[9] For example, Angrist (1990) used the Vietnam-era draft lottery to create an instrument for military service, allowing an estimate of the effect of such service on later wages controlling for selection bias into the military.

sample of firms that received Small Business Innovation Research (SBIR) grants from various Federal agencies over a number of years, and a set of similar firms that did not. For each firm (whether funded or not), he calculates a weighted-average available budget in each year using firm-agency weights derived from the SIC of the firms and the SIC distribution of firms funded by each agency. He shows that using this effective budget as an instrument for SBIR funding greatly reduces the apparent impact of SBIR funding on performance in an equation analogous to Equation (1).[10]

I am not aware of any attempts to use affirmative-action related variables as instruments in this kind of analysis (although the possibility was noted by Arora and Gambordella, 1998). If being female, minority, from a heartland state, or from a 'second tier' institution increases the probability of funding, conditional on expected performance, then these attributes are possible instruments. Similarly, if agency decisionmakers attempt to balance their funded portfolios across technical subfields, then proposals from subfields that are underrepresented in the proposal pool have a higher conditional likelihood of funding, so that subfield dummies might also be possible instruments.[11] The question regarding this approach is whether these considerations have enough impact on the overall selection probability to constitute adequately powerful instruments for selection.

III. DESIGNING RESEARCH GRANT PROGRAMS TO FACILITATE SUBSEQUENT EVALUATION

The above discussion suggests that there are empirical strategies for measuring treatment effects despite the selection problem. This section considers whether changes in the grant decision process could improve the ability of subsequent evaluations to provide reliable measurement of the treatment effect.

A. Randomization

The conceptually straightforward way to solve the selection problem is to run an experiment. That is, since the 'problem' is that the likelihood of treatment is correlated with expected performance, the simplest solution

[10] Lichtenberg (1988) used a similar instrument for defense procurement funding in a study of its impact on firm-level R&D.

[11] A technical problem with using technical subfields as instruments is that, even if true expected performance is not a function of subfield, *measured* performance might be. That is, if IO economists publish more (or fewer) papers than labor economists, controlling for quality, then you would want to have subfield dummies in estimating Equation (1) using papers as the measure of output or performance. This would invalidate the subfield dummies as instruments for selection.

is to make the probability of selection conditional on $\alpha_i + \omega_{it}$ the same as the unconditional probability. To do this the agency would identify a group of 'potential grantees' and randomly award grants within this group, meaning that the probability of receiving the grant would be the same for all members of the group.[12] If this were done, then a later evaluation could estimate the average treatment effect for the group by estimating Equation (1) on all group members, funded or not. Because the probability of having been treated is the same for all group members, the estimated treatment effect is not subject to selectivity bias.[13]

The group of *potential* awardees could be identified in a variety of different ways. In particular, the agency could eliminate from the application pool a subset of clearly inferior proposals, and then award funding randomly only to the remainder. This prior 'selection' into the *potentially* funded group would not introduce any bias into the estimate of the average treatment effect for the treated group. That is, the estimated effect could not be extrapolated to the group of applicants who were screened out in advance, but it would be valid, as an average, for the entire group that was given a chance of funding.

Multiple subgroups could be randomized in this way. For example, the set of applicants could be divided into 3 groups: 'high priority', 'marginal', and 'rejects', with the probability of receiving funding higher for the 'high priority' group than the 'marginal' group, and no funds awarded to the 'rejects'. The estimation of Equation (1) could then be carried out separately for the two groups, yielding distinct estimates of the average treatment effect for the 'high priority' and 'marginal' groups. If an overall average treatment effect for the treated groups was desired, a weighted average of the two estimated effects could be calculated.

The obvious political and ethical concern about this approach is that some 'high priority' proposals are left unfunded.[14] Of course, one could

[12] Randomization might also be desirable independent of evaluation objectives. Brezis (2000) presents a model in which the selection agency introduces randomization to make sure that a larger fraction of radical proposals—which are assumed to be undervalued by the review process—are funded.

[13] There is, of course, selection by the applicants into the public process. This means that even randomly awarded grantees could not be compared to entities that did not apply for funding. But the treatment effect could be estimated from data on grantees and rejected applicants. Of course, the estimated effect could not then be extrapolated to the population of non-applicants. Further, the running of the experiment might change the applicant pool, so that the observed treatment effect might not be what occurs when the applications are generated in the absence of the experiment.

[14] I remain personally puzzled as to why it is okay to randomize when people's lives are at stake (drug trials), but not when research money is at stake. When I put this question to an agency official who was quite hostile to randomization, he pointed out to

make the actual probability of funding in the high priority group quite high. The only limitation on how high this probability could be made is that the absolute number of unfunded proposals must be great enough to produce estimates of Equation (1) with adequate precision.[15] In the limit, one could make the probability of funding in the 'high priority' group unity, and randomize only within the marginal group. The consequence would be that one would have an estimate of the average effect of treatment for marginal applicants, but no estimate of the treatment effect for the high priority applicants. Since 'high priority' ought to mean that the agency expects the treatment effect to be large, this estimate of the treatment effect for the marginal group should be an underestimate of the overall average treatment effect for the treated group. One would then have bounds on the overall average treatment effect, running from the estimate from the randomized marginal group, up to the selection-biased estimate derived from Equation (1) in the full sample. It is unclear, however, whether the value of the lower bound derived from the randomized marginal group is worth the political pain of introducing randomization.

B. The Regression-Discontinuity Design

Thus the problem with using randomization to eliminate selectivity bias is that one must either deny funding to high-priority proposals, or else accept that one cannot produce an unbiased estimate of the treatment effect for such proposals. I believe that an alternative approach, based on the regression-discontinuity design, offers a more attractive balance between political feasibility and statistical outcome.[16]

The regression-discontinuity ('RD') design was introduced by Thistlethwaite and Campbell (1960); good overviews appear in Campbell (1984) and Angrist and Lavy (1999). The RD technique utilizes a discontinuity in the probability of selection that occurs at a particular 'threshold' with respect to some index of 'quality' to identify the treatment effect separately from the impact of quality. To make this concrete in the current context, imagine that the review panel or

me that once a drug has been demonstrated to be effective, random trials are ended and all patients are given the drug. When I commented that the relevance of this to research funding depends on an assumption that the efficacy of funding has been demonstrated, he responded 'of course'.

[15] It is also unclear whether it would be perceived as less 'unfair' to have a small fraction of deserving proposals unfunded than to have a larger fraction unfunded.

[16] The only previous application of the RD design to a research grant program that I have been able to identify is Carter, Winkler and Biddle (1987), who evaluated the NIH Research Career Development Award ('RCDA'). They found that RCDA recipients had significantly greater research output than non-recipients, but that, after controlling for the selection effect, there was no detectable effect of the RCDA 'treatment' itself.

committee ranks each applicant from best to worst, and records each applicant's ranking.[17] The selection official then 'draws a line' through this ranking, indicating that, if quality were the only issue, proposals above the line would be funded and those below would not. The location of this threshold is recorded. The official then makes the actual award decisions; these can deviate from those implied by the threshold, so long as the deviations are for affirmative action or other non-expected-performance-related reasons.

Equation (1) could later be estimated on data from funded and unfunded projects, using the quality ranking as one of the X's, and using an indicator variable for ranking above the threshold as an instrument for selection. The quality ranking controls, by construction, for anything that the selection process knows about α_i and ω_{it}. And, conditional on this ranking variable, the threshold-indicator variable is a valid instrument for selection. In effect, we have used the known discontinuity in selection probability at the threshold to create an exclusion restriction based on functional form: the threshold indicator variable is simply a non-linear function of the quality rank. But this functional form assumption is founded in the selection process itself, not imposed on the distribution of an unobserved latent variable.[18]

Without further assumptions, this approach only identifies the treatment effect *at the threshold quality level*. In this, it is comparable in the information it generates to the use of randomization only for a group of 'marginal' applicants. Why this is true can be seen from Figure 1. The figure plots a hypothetical relationship between selection rank and research output. As drawn, output increases with rank, treatment increases output, and the treatment effect increases with rank. That is, the best proposals have both higher expected output without government support, and also a larger increase if they are supported. 'Most' of the proposals above the threshold were funded, and most of those below were not, but the figure shows a few above-threshold ones that were not funded and vice versa, to reflect the idea that random deviations from the threshold may have occurred.

[17] It is not actually necessary that all proposals be ranked. In particular, if there is a group of clear rejects at the bottom, they need not be ranked. Also, there could be ties, i.e., groups of applicants judged to be equally meritorious.

[18] If data are available regarding the attributes that form the basis of non-expected-performance-related deviations from the funding decisions implied by the threshold (e.g., gender, race, proposing institution, etc.), then additional instruments related to these characteristics would, in principle, improve the first-stage fit between the instruments and the selection dummy, and hence increase the power of the procedure.

The smooth *dashed* line represents a regression estimate of the relationship between rank and output for the funded proposals, and the smooth *solid* line an estimate of this relationship for the unfunded proposals. Both of these are identified from the observed data, and, therefore, so is the gap between them at the threshold. Of course, this is only a lower bound to the average treatment effect for the treated group. Now, if we are willing to extrapolate the solid trend line beyond the threshold—into a range in which we have very few observations—then we can estimate the treatment effect for each of the funded proposals, and its overall average.[19] This requires assumptions about the functional form of the rank/output relationship that are presumed to hold over the entire rank range.[20]

In summary then, both 'randomization at the margin' and the regression-discontinuity design can, at least in principle, provide a basis for unbiased estimates of the treatment effect for the marginal proposals. This is, however, only a lower bound for the average treatment effect. To do better through randomization requires denying funding to a (statistically) significant number of highly ranked proposals. To do better via the regression-discontinuity design requires willingness to rely on functional form assumptions for the relationship between selection rank and output, in order to 'predict' the unobserved expected performance of the best proposals had they not been funded.

IV. THE 'ADDITIONALITY' QUESTION AND THE PRODUCTIVITY OF PUBLIC RESEARCH FUNDING

Much discussion regarding the performance of public research support programs (particularly those that support firms) has focused not on research output, but on the related question of 'additionality'—the extent to which public grants lead to an increase in overall research expenditure by the funded firms. Results on this question are mixed.[21] Wallsten (2000) found that the SBIR program 'crowds out' the firm's own research spending approximately dollar-for-dollar, reversing the finding of Lerner (1999) for this same program. Branstetter and Sakakibara (2000) found

[19] A similar extrapolation would be necessary to estimate what the treatment effect *would have been* if the rejects had been funded. But this is less interesting.

[20] Note that affirmative action or other deviations from the threshold rule help in this respect. This is not surprising, because they generate, in effect, a small amount of 'random' data that provide additional identification.

[21] David, Hall and Toole (2000) survey the econometric evidence on this issue. They find that a plurality of studies at the firm level finds 'crowding out' effects, while studies at higher levels of aggregation more often find 'crowding in'. They also note that difficulties of econometric interpretation make it difficult to draw robust conclusions from these studies.

that Japanese funding of research consortia increased the R&D of the participating firms. Lach (2000) found that research support of commercial firms in Israel increased the firms' total R&D expenditure by $1.41 for every dollar of public research expenditure.

Measuring the effect of government support on total research expenditure is, of course, just as subject to the selection bias problem as measuring the effect on research output. Those firms (or academic researchers) funded by the government are likely to be those with the best ideas, meaning that they will have more incentive to spend their own money, and more ability to garner support from third parties, than those that are not funded. Any regression analysis that compares the research expenditure of supported firms to those that are not supported has to deal with all of the problems discussed above. As emphasized by David and Hall (2000), however, the additionality question is also plagued by confusion regarding the underlying model of how firms, agencies, and other parties make research spending decisions.

Wallsten (2000) describes the straightforward argument for crowding out. If there are short-run diminishing returns to R&D, and the firm spends its own money up to the point where the expected marginal return is equal to the cost of funds, infusion of funds by the government will cause the firm to reduce its own expenditure dollar for dollar, so that the total funding (and hence the expected marginal product) remains the same. Some grant agencies, however, require cost-sharing or co-funding of research proposals by the proposing firm. Depending on how this is implemented, it could be interpreted to mean that (if selected) the firm gets additional public funding for every additional dollar of funding it provides itself or from other sources. If the co-funding rules work this way, the effect is to reduce the marginal cost of research to the firm. A profit-maximizing firm facing a downward sloping marginal research returns schedule will always increase *total* expenditure when the marginal cost falls, precluding the dollar-for-dollar crowding out result. The amount by which total R&D increases with public funds would depend on how rapidly marginal productivity diminishes, suggesting some unknown degree of partial crowding out.

Another reason that the Wallsten argument for crowding out may not apply is that the funding agency is often picking certain projects to fund, presumably because these particular projects are believed to have large social returns. If such projects are far down on the private marginal returns schedule, then they would not be undertaken by an unsubsidized firm, but may be undertaken if the government is willing to fund them. In this case, total R&D expenditure increases and there may be no crowding out.

If the only reason for public support of private research were a belief that spillovers cause a gap between the private and social rates of return, then one could think about this issue in terms of the relative size of the spillover gap and the crowding out effect. In many contexts, however, we think that the issue is not limited to profit-maximizing firms choosing the wrong point on the marginal product of research schedule. In some cases this is because the researchers are not in a for-profit setting. Even with respect to firms, however, we believe that there are constraints on the financing of R&D.

The presence of financing constraints creates the possibility of 'crowding in' (Diamond, 1998) of private research funding. Screening of proposals for likely success is a costly and uncertain process. The public funding decision represents certification of a proposal as 'high quality'. Non-public sources of funding may free-ride on the public review process, or, even if they make their own assessments, know that their assessment is uncertain and be influenced by the assessment of government experts. This 'certification' or 'halo' effect is believed by research grant agencies in the U.S. to be an important factor in increasing the total research spending of grant recipients.

Once one recognizes these complexities, the relationship of the additionality question to the underlying public policy issues becomes ambiguous. I presume that the underlying policy motivation for GPRA is that, in some sense or on some level, we want to perform social benefit-cost analysis of these programs. We want to know the marginal (or more realistically, the average) social product of these public expenditures. Additionality, in the sense discussed here, is neither necessary nor sufficient for public funding to yield a positive social product. The consequences of either crowding out or crowding in depend on the opportunity cost of the alternative funds. For example, if alternative sources of funding are in fixed supply—for example, the money available for research from non-profit foundations—the policy implications of crowding out are unclear. It could be that public funding 'crowds out' foundation funding of government-supported researchers, pushing the foundation funding over to other (productive) researchers whom we don't observe. In this case, we might observe no impact of public support on the supported researchers, yet the social product could still be large. [22] Conversely, a 'halo' effect could simply induce a zero-sum redistribution

[22] An analogous argument could be made with respect to commercial support programs and venture capital finance, though the argument for a fixed overall supply of venture capital for funding of research seems more questionable.

of available non-governmental funding. [23] Thus the extent to which crowding out limits or crowding in enhances the social productivity of public research funding depends on the elasticity of supply of alternative sources of funds.

Unfortunately, similar reasoning leads to similar conclusions regarding what can be learned from estimation of Equation (1) without regard to additionality, i.e., 'solving out' the effects of government support on total funding and measuring the overall impact of the public funding decision on research output. A positive average treatment effect in Equation (1) is not sufficient for inferring a positive social product of the government funding; it could reflect totally unproductive use of the government funds combined with a 'halo' effect that pulls in (productive) funds from elsewhere. If those other funds are in fixed supply, this is not socially useful. And a positive effect on research output is not necessary for a research support program to be productive, because a lack of treatment effect could reflect total crowding out, which, as discussed above, could still yield positive social product if other research funding sources are in fixed supply.

Thus the crucial question is not the extent of additionality *per se*, but rather how, in the presence of the possibility of crowding out or crowding in, can one measure the social product of the public research funding itself. To do this, we need to modify Equation (1) to incorporate explicitly both the productivity of public research support, and its impact on non-governmental spending:

$$Y = \lambda_1 X + \gamma^P P + \gamma^G G + \alpha + \mu + \omega + \varepsilon \qquad (2a)$$

$$P = \lambda_2 Z + \delta_i G + \alpha + \mu + \omega + \varepsilon \qquad (2b)$$

where P and G represent private and government research expenditure, respectively, Z is a vector of characteristics that affect the level of private funding, and the i and t subscripts have been suppressed for convenience.[24] The 'treatment effect' from government research support has been separated into two pieces: γ^G reflects the direct effect, while the sign of δ tells whether we have crowding in or out. Identification of this

[23] Diamond (1998) finds evidence of 'crowding in' in aggregate totals for research support. This suggests that the supply of non-governmental research support is not fixed.

[24] Different functional forms that capture the same ideas are also possible. For example, one might want to do $\gamma \log(P + \gamma^G G)$ rather than an additive form. Also, one might think that, once the public expenditure is present, there should be no effect of selection *per se*, other than through a possible halo effect that is captured in the second equation.

model raises all of the issues discussed above, plus the need for instruments for private expenditure.[25]

The welfare-analytic implications of γ^G are straightforward; it captures the average productivity of the funds expended by the government. But, as discussed above, the welfare consequences of δ depend on assumptions about the larger financing system with which the government agency interacts. This takes us back to where we started, which was to note that measuring the direct impacts of public funding on the funded entities is only a small piece of the overall evaluation problem.

V. CONCLUSION (RUMINATIONS ON POLITICAL ECONOMY)

The pressure on public research agencies to engage in systematic evaluation of their programs is likely to continue to grow. While much can be learned from what I have called 'after-the-fact' evaluations, the reliability and 'believability' of these results in the face of presumed selection bias could be increased by building evaluation needs into the grant process. I and others have previously harped on randomization as the 'gold standard' for program evaluation (Jaffe, 1998). I now believe that the regression-discontinuity design offers a better tradeoff between statistical benefits and resistance to implementation. In particular, randomization at the margin, which seems like something that one might be able to 'sell' both to a funding agency and to its constituents, has the major drawback of providing only a lower bound on the overall effect of agency funding on research outputs. While the RD design requires functional form assumptions to do better than this, we are frequently willing to make such assumptions.[26]

I believe that the use of this technique would be good social science. Whether it would be good for the agencies in question—or for public policy more generally—is a much harder question. As social scientists we are interested in the parameters of Equation (2) even though it is difficult to know exactly how they relate to benefit cost analysis. We are comfortable examining this tiny piece of a very complicated puzzle, ignoring as we do so that our output measures are only proxies for what

[25] One might also argue that G is endogenous, if the size of the grant (and not just the selection decision) is related to the unobservables or to P.

[26] One might ask whether functional form assumptions could be used in conjunction with randomization at the margin in order to produce an estimate of the overall treatment effect. This would require recording of the selection ranking, in order to estimate the relationship between ranking and output. Once one has recorded the ranking, then you essentially have the RD design. Given the RD design, there is relatively little benefit to randomizing near the threshold.

we care about, that we are not looking at the spillovers that are perhaps the true reason for these programs, that we have a hard time capturing the long-term effects of funding on research careers, and that we are not measuring the ‘general equilibrium’ interactions between the funded researchers and the rest of the system. We also understand that failing to reject a null hypothesis is not the same as showing the null to be true. I am, of course, aware that the political process may ignore these subtleties and misuse research findings no matter how many caveats appear in the papers reporting those findings. I therefore will not pretend to know the answer regarding the overall social benefit-cost ratio of undertaking these kinds of evaluations.

References

Angrist, J. (1990), 'Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence from Social Security Administrative Records', *American Economic Review*, **80**, 313-335.

Angrist, J., and Lavy, V. (1999), 'Using Maimonides' Rule to Estimate the Effect of Class Size on Scholastic Achievement', *Quarterly Journal of Economics*, **114**, 533-576.

Arora, A., and Gambardella, A. (1998), 'The Impact of NSF Support for Basic Research in Economics', mimeo, Carnegie-Mellon University.

Branstetter, L., and Sakakibara, M. (2000), 'When Do Research Consortia Work Well and Why? Evidence from Japanese Panel Data', National Bureau of Economic Research Working Paper No. 7972.

Brezis, E. S. (2000), 'Randomization: An Optimal Mechanism for the Evaluation of R&D Projects', mimeo, Bar-Ilan University.

Brown, M., Curlee, T. R., and Elliott, S. R. (1995), 'Evaluating Technology Innovation Programs: the Use of Comparison Groups to Identify Impacts', *Research Policy*, **24**, 669-684.

Campbell, D. (1984), *Research Design for Program Evaluation*, Beverly Hills, Sage Publications.

Carter, G. M., Winkler, J. D., and Biddle, A. K. (1987), *An Evaluation of the NIH Research Career Development Award*, R-3568-NIH, Santa Monica, The Rand Corporation.

Cozzens, S. E. (1999), 'Are New Accountability Rules Bad for Science?' *Issues in Science and Technology*, Summer.

David, P. A., and Hall, B. H. (2000), 'Heart of Darkness: Modeling Public-Private Funding Interactions Inside the Black Box', National Bureau of Economic Research Working Paper No. 7538.

David, P. A., Hall, B. H., and Toole, A. A. (2000), 'Is Public R&D a Complement or a Substitute for Private R&D? A Review of the Econometric Evidence', *Research Policy*, **29**, 497-529.

Diamond, A. M. (1998), 'Does Federal Funding "Crowd In" Private Funding of Science', *Contemporary Economic Policy*, 423-431.

Georghiou, L., and Roessner, D. (2000), 'Evaluating Technology Programs: Tools and Methods', *Research Policy*, **29**, 657-678.

Griliches, Z. (1992), 'The Search for R&D Spillovers', *Scandinavian Journal of Economics*, **94**, S29-S47.

Guellec, D., and van Pottelsberghe, B. (2000), 'The Impact of Public R&D Expenditure on Business R&D', DSTI Working Paper, Paris, Organization for Economic Cooperation and Development.

Hall, B., and Van Reenen, J. (2000), 'How Effective Are Fiscal Incentives for R&D? A Review of the Evidence', *Research Policy*, **29**:449-469.

Heckman, J. H., Ichimura, H., Smith, J., and Todd, P. (1998), 'Characterizing Selection Bias Using Experimental Data', *Econometrica*, **66**, 1017-1098.

Jaffe, A. B. (1998), 'Measurement Issues', in L.M. Branscomb and J.H. Keller (eds.), *Investing in Innovation*, Cambridge, MIT Press.

Jones, C., and Williams, J. (1998), 'Measuring the Social Return to R&D', *Quarterly Journal of Economics*, **113**, 1119-1135.

Kauko, K. (1996), 'Effectiveness of R&D Subsidies—A Sceptical Note on the Empirical Literature', *Research Policy*, **25**, 321-323.

Klette, T., Møen, J., and Griliches, Z. (2000), 'Do Subsidies to Commercial R&D Reduce Market Failures? Microeconometric Evaluation Studies', *Research Policy*, **29**, 471-495.

Lach, S. (2000), 'Do R&D Subsidies Stimulate or Displace Private R&D? Evidence from Israel', National Bureau of Economic Research Working Paper No. 7943.

Lerner, J. (1999), 'The Government as Venture Capitalist: The Long-Run Impact of the SBIR Program', *Journal of Business*, **72**, 285-318.

Lichtenberg, F. (1988), 'The Private R&D Investment Response to Federal Design and Technical Competitions', *American Economic Review*, **36**, 97-104.

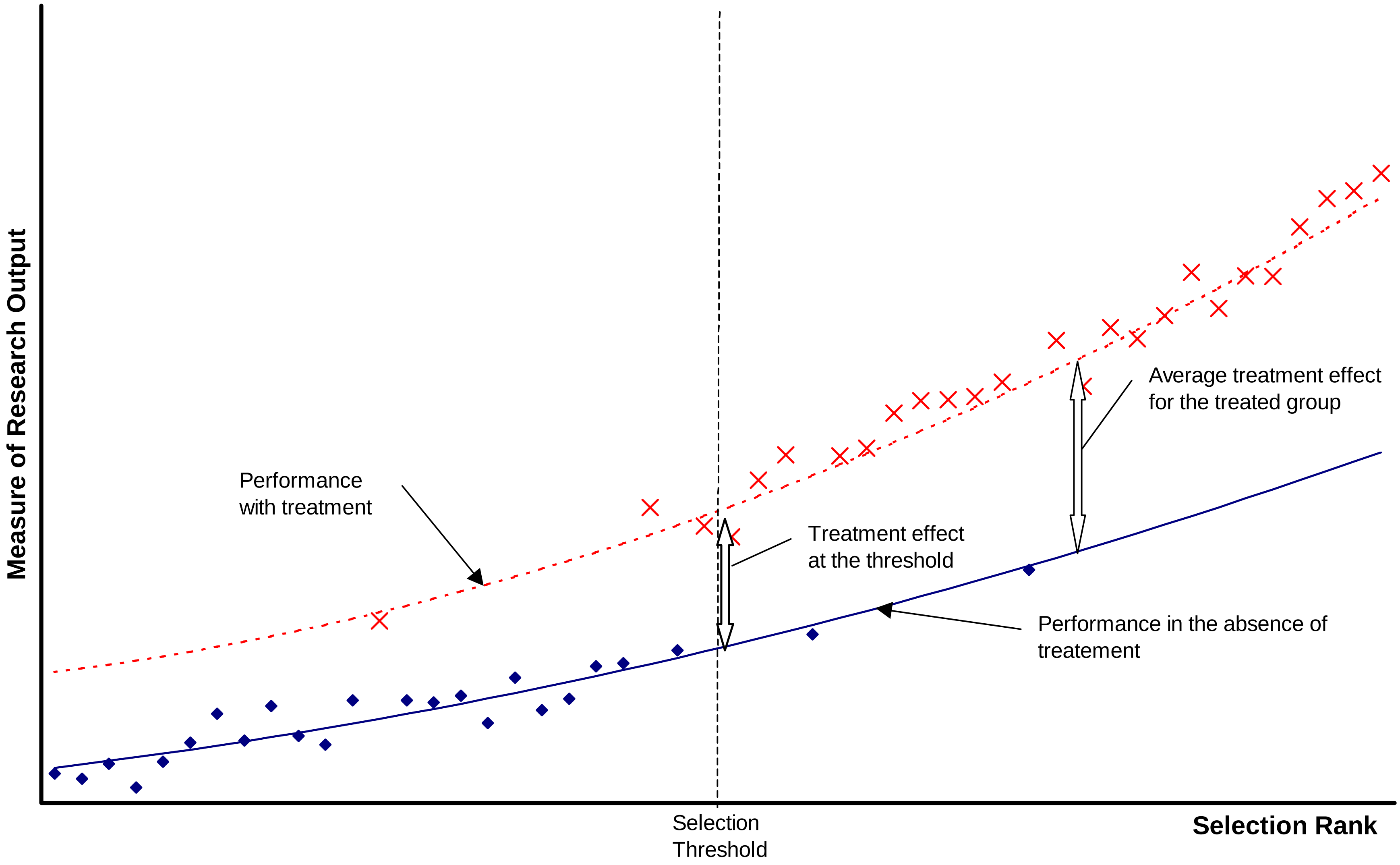
Popper, S. W. (1999), 'Policy Perspectives on Measuring the Economic and Social Benefits of Fundamental Science', MR-1130-STPI, Washington DC, Rand Corporation Science and Technology Policy Institute.

Thistlethwaite, D., and Campbell, D. (1960), 'Regression-Discontinuity Analysis: An Alternative to the Ex Post Facto Experiment', *Journal of Educational Psychology*, **51**, 309-317.

Wallsten, S. (2000), 'The Effects of Government-Industry R&D Programs on Private R&D: The Case of the Small Business Innovation Research Program', *Rand Journal of Economics*, **21**, 82-100.

Yager, L., and Schmidt, R. (1997), *The Advanced Technology Program: A Case Study in Federal Technology Policy*, Washington DC, AEI Press.

Figure 1
Analysis of Hypothetical Data from a Regression-Discontinuity Design


Measure of Research Output
Performance
with treatment
Treatment effect
at the threshold
Average treatment effect
for the treated group
Performance in the absence of
treatement
Selection
Threshold
Selection Rank