Policy Reform, 2002, Vol. 5(4), pp. 251–269

It Pays to be Ignorant: A Simple Political Economy of Rigorous Program Evaluation

LANT PRITCHETT*

John F. Kennedy School of Government, Harvard University

(Received 25 October 2002; In final form 14 March 2003)

This paper attempts to explain the scarcity of rigorous evaluations of public policy. I build a positive model to explain the "stylized fact" that there is under investment in the creation of reliable empirical knowledge about the impacts of public sector actions. The model shows how "advocates" of particular issues or solutions – the public action equivalent of entrepreneurs – have incentives to under invest in knowledge creation because having credible estimates of the impact of their preferred program may undermine their ability to mobilize political (budgetary) support.

Key words: Program evaluation; Bureaucracy; Issue advocates

JEL: D23, D73, H43, O1

1 INTRODUCTION

Nor certitude, nor peace, nor help for pain;
And here we are as on a darkling plain
Swept with confused alarms of struggle and flight
Where ignorant armies clash by night
Matthew Arnold, Dover Beach
It pays to be ignorant, to be dumb, to be dense, to be ignorant . . .
TV game show It Pays to be Ignorant (1949) theme song

This paper was motivated by my dozen years in the World Bank, a large, international, quasi-public, bureaucracy whose objective was "development" and whose instrument was providing loans to developing country governments. The organization's lending activities have spanned the range: from dam construction to family planning to micro credit to steel mills to "social funds" to macroeconomic stabilization to land reform. The World Bank is for the most part staffed by internationally recruited, highly trained, well-meaning, and experienced professionals and is arguably the premiere development institution. And yet, nearly all World Bank discussions of policies or project design had the character of "ignorant armies clashing by night" – there was heated debate amongst advocates of various activities but very rarely any firm evidence presented and considered about the likely impact of the proposed actions. Certainly in my experience there was never any definitive evidence that would inform decisions of funding one broad set of activities versus another (*e.g.* basic education versus road

* John F. Kennedy School of Government, Harvard University, 79 John F. Kennedy Street, Littauer-313, Cambridge, MA 02138, USA; Tel./Fax: (617) 495-1100; E-mail: lant_pritchett@ksg.harvard.edu

ISSN 1384-1289 print; ISSN 1477-2736 online
DOI: 10.1080/1384128032000096832

construction or vaccinations versus macroeconomic reform) or even funding one instrument versus another (*e.g.* vaccinations versus public education about hygiene to improve health, textbook reform versus teacher training to improve educational quality). How can this combination of brilliant well-meaning people and ignorant organization be a stable equilibrium?

In the United States no one can market a prescription medicine for male pattern baldness without evidence it is "safe and effective". The accepted regulatory standard evidence of safety and effectiveness is a controlled, randomized, double blind, evaluation. Yet the non-profit "market" is flooded with a continual new stream of proposed programs and interventions. Few public sector actions, even those of tremendous importance, are ever evaluated to the standard required of even the most trivial medicine. To take just one example, in the United States there is a huge and continuing debate over the importance of smaller class sizes for academic performance in primary and secondary education. One side of the debate points to the fact that per pupil expenditures in public schools have doubled while test scores have changed very little and to many studies which find no effect of class size to argue that it is plausible that hundred of billions of dollars of educational resources have been misallocated. The other side of the argument suggests that smaller class sizes are associated with stronger performance. The point is not that one side is obviously right and the other obviously wrong – the point is that brilliant, well meaning people can legitimately disagree on so fundamental a question as class size impacts on educational quality – yet there is no similar debate on the efficacy of treatments for male pattern baldness.

There is no question that estimating the impact of public activities on outcomes is harder than the science of hair loss. But it is not impossible. The key problem with evaluating the impact of any public program is that it depends not just on facts but also on a counter-factual. Even if what happened to program participants before and after the program is well documented, to know the impact of the program one has to know *what would have happened* to the program participants in the absence of the program. For very good reasons the "gold standard" of program evaluation is controlled experiment in which participation (or access to) the interventions is randomized.

Since this is the "gold standard" the scarcity of this gold constitutes a major puzzle. That the exceptions are so well known proves the rule:

- Head Start is a program that consistently receives favorable attention, much of which is based on a randomized evaluation (actually not of Head Start but of a more intensive program) involving only 123 children.
- In a search on the Econlit there are 29 references to the Job Training and Partnership Act (JPTA). Why? Not because it was ever a particularly large or important federal program – in the 1990s its budget was around US $1.6 billion – but Title II of the JPTA provided for the largest randomized evaluation of training and hence analysts use the data over and over again – even though the program itself was ended.
- Family planning programs have existed and been promoted the world over for thirty years, motivated and supported at least in part by the belief they lower fertility. Yet there is exactly *one* reliable randomized experiment (carried out in the Matlab region of Bangladesh) that plausibly demonstrates a fertility reduction impact of increased supply of contraception (Phillips *et al.*, 1988).
- The optimal structure of insurance is an enormous policy issue, and yet there is one randomized experiment that attempted to demonstrate the impact of alternative health insurance options in the United States (Manning *et al.*, 1987).
- "Primary health care" is a mantra widely endorsed (if not implemented) at least since the Alma Ata conference in 1978. But there is not a single rigorous evaluation

demonstrating the health gains from a "primary health care" as a public policy.[1] (Filmer *et al.*, 2000).

- There are debates about the relative magnitude and importance of various educational inputs – smaller class sizes? More books? Better facilities? Yet in the developing world there are very few controlled experiments in education – and even fewer that are successfully completed and produce reliable results (Kremer *et al.*, 1997).

As the examples from the US show, this is not just a "development" problem – but the scarcity of rigorous evaluation is particularly striking in development assistance in which donors finance discrete "projects" – nearly always with a "monitoring and evaluation" component. Yet the World Bank has lent a cumulative 100 billion dollars in and only a handful have produced compelling evidence about impacts from a reliable, randomized evaluation.[2]

This dearth of knowledge is sufficiently striking as to deserve explanation and common explanations casually proposed – ethical barriers, costs, and feasibility – are not sufficient (see below). Here I construct a simple analytical model that explains both why randomized evaluations are infrequent and when evaluations can be expected.

The model focuses on the decisions faced by advocates of particular public (or non-profit) sector activities who must mobilize funds. "Advocates" are the entrepreneurs of the public and non-profit private sector. Just as with their profit seeking counter-parts in the private sector, advocates are the men and women with the passion, concern, and vision to make new things happen. Programs are typically proposed, supported, and implemented by "advocates" – people with strong beliefs in either the importance of a particular *issue* – the problem the public program addresses (crime, drugs, unemployment, malnutrition) or a particular *instrument* – the particular solution to a problem (community policing, "12 step" therapy, job training, micro-nutrient supplementation, micro-credit). Without advocates pushing to address issues and for innovative instruments – new programs, policies, treatments, the public and not-for-profit sectors would be stagnant. Yet this same dynamism in the face of the intrinsic uncertainties of a complex world can lead to tensions between action and knowledge.

At the core of the model are two assumptions about evaluations and about political support for particular programs. The first assumption is that a randomized evaluation is impossible without the cooperation of the advocates responsible for program implementation so that evaluations can only happen if advocates see them in their best interest.

The second assumption is that advocates are more altruistic and care more about outcomes in their specific issue than does the general public. Given this concern for outcomes they want to pursue the most effective instrument and, at any given level of the efficacy of the use of resources (outcome gain per dollar), they want a larger budget. If the budget is politically determined advocates view the problem of evaluation in a dual light. On the positive side evaluations potentially help improve program efficacy so they get more bang for the buck. But evaluations have a potential downside if they reduce political support for a larger budget for their program so they get fewer bucks.

In this model advocates may choose ignorance over public knowledge of true program efficacy. They are better off if the voting public does not know the "true" benefits even if it means they too must operate somewhat in the dark.

[1]This is about PHC as a policy as opposed to the medical gains from specific health interventions like vaccinations. One of the persistent confusions in the literature on health policy is precisely between the effectiveness of specific actions as medical "interventions" and the impact of a public sector "intervention" (Hammer and Pritchett, 2001). It is perfectly possible that a certain therapy can be medically effective and yet a policy of public provision of that therapy have no impact on health status of the population.

[2]Every World Bank loan has an "evaluation" but many of these do not involve any attempt to estimate the magnitude of impact on specific outcomes and rarely is there any estimate of the "without the project" counter-factual. This is not a criticism of the World Bank as other donor agencies are even less advanced in their evaluation methods.

I build this model in three separate steps.

In section I take the point of view of "true believers" – advocates who believe they know both the important *issue* and the correct *instrument* to address that issue. In this case evaluations are rarely useful – and hence will be rare – as evaluations do not increase advocates' knowledge about program efficacy (as they already know) and yet they risk cutting political support.

Second, in organizations which are coalitions of advocates for the same *issue* (*e.g.* biodiversity, maternal health, basic education) but in which there is uncertainty about the efficacy of various *instruments*, evaluations are needed to improve efficacy of the utilization of organizational resources but the outcomes of these evaluations could undermine budgetary support for the organization.

Third, the most complex case is one where organizations are coalitions of advocates around a broad objective (*e.g.* development, global health) that encompass several *issues* and there is a problem of deciding which is the most important issue and which is the best *instrument* to address the issue. The political constraints imposed by the less altruistic part of the public forces this coalition of advocates into an uneasy coalition that supports ignorance in which efficacy of budget use is sacrificed to maintain budget size.

2 THE POLITICAL ECONOMY MODEL

2.1 True Believers: Single Objective Single Instrument Advocates

In the model that I will develop there are four actors, advocates or coalitions of advocates and three groups of "the public". In the simplest version of the model advocates are single objective, single instrument advocates. That is the advocates not only know what is an important problem they know the best solution (*e.g.* population reduction is the objective and family planning is the instrument, raise welfare recipient's wages is the objective and job training is the instrument, etc.). Since these advocates believe they know the efficacy of the instrument their only goal is to increase total expenditures on their preferred instrument.

2.2 Savvy Altruists

Assume that advocates:

(a) Believe they know the true value of program efficacy,
(b) believe with certainty that a rigorous evaluation will reveal that true value,
(c) have linear utility in program expenditures net of first period costs.[3]

Advocates take all the actions in the first period. They are the ones that initiate, administer, promote budget expansion and, perhaps, evaluate programs. We assume a rigorous evaluation is impossible without advocate's cooperation – a binding requirement of rigorous evaluation is only possible in this model if it creates incentives such that advocates choose to evaluate.[4]

[3]I ignore discounting of the first period costs because it affects C and S equally so will not tip the decision rule.

[4]This assumption can be justified in three ways. First, experience in the world suggests that far more rigorous impact evaluations are promised than begun, more begun than finished. From my observation evaluation tends to be always low on the priority of the program initiators and advocates and hence just never gets done. Second, politically it would be almost impossible for a program evaluation finding "no impact" to be credible if it were undertaken by groups who were not in some sense program advocates. Third, there almost certainly could be more fundamental assumptions made about the observability of effort in which case it might be impossible to credibly contract with advocates to undertake actions not in their best interest.

Advocates have only two options: "Pilot and Persuade" or "Rigorous Evaluation". If "Pilot and Persuade" is chosen then in the first period advocates "pilot" – spend just enough money on implementation to demonstrate to a reasonable person that the program is not physically impossible – and "persuade" – spend amount (S) on the dissemination of non-rigorous claims about the efficacy of the program. If "rigorous evaluation" is chosen, the advocates implement and spend the cost of a randomized evaluation (C) in the first period.

In this case the conditions for evaluation reduce to the question of whether the equilibrium chosen level of program expenditure level (E(·)) net of costs of either persuasion or evaluation is higher with or without a rigorous evaluation (Eq. (1)):

$$\text{Do evaluation if: } E^K(\alpha_{\text{true}}, \alpha_I^C) - C > E(S^*) - S^*(\alpha_I^0, \beta_I^C) \tag{1}$$

Which of these options is in the best interest of the advocates depends on how the level of expenditures is determined. For this I have a simple voting model with three groups.

2.3 The Voting Public

After the advocates take their actions the "public" votes over the level of the budget for the program. I assume the public is classified into three groups, core supporters, hard headed and the middle group. The utility of these groups differ in two respects, the extent to which program gains enter their utility (γ) and how effective a dollar's worth of program expenditure will be in producing program gains (α).

The program efficacy parameter depends on many factors. Higher overhead costs lead to lower beneficiary impact per dollar and hence lower program efficacy. Better targeting to the poor for given dollar amount of benefits could raise α. Program efficacy beliefs could also reflect "specific altruism" so that group "i" is not indifferent between a program that delivers a dollars worth of benefits in cash (spent as the recipient chooses) to one that delivers a dollar's worth of some good group "i" believes has particular merit (food for children, medical care, family planning).

$$U_i(X, E) = \gamma_i \alpha_i(\cdot) + (1 - \gamma_i) U(X_i) \tag{2}$$

All three groups have exactly the same decision rule: they will vote for an additional dollar of program funding if their own utility gain from benefiting others via that program[5] exceeds the utility loss from foregone own consumption.[6] Hence the decision rule is the same for all groups:

$$\text{Support if: } \alpha_i > \alpha_i^{\text{Critical}} = \left(\frac{\partial U}{\partial X}\right) * \left(\frac{1 - \gamma_i}{\gamma_i}\right) \tag{3}$$

We assume that each group forms beliefs about program efficacy in the same way. If there has been evaluation then all believe the efficacy is the outcome of the evaluation (this assumes the evaluation measures all the relevant dimensions of efficacy).

$$\text{If evaluation, beliefs are: } \alpha_i = \alpha^{\text{RE}} \tag{4}$$

[5]There is a more cynical, and hence more realistic, version in which "core supporters" benefit more or less from the program spending itself (*e.g.* recipients of social security, farmers lobbying for subsidized sales of food to developing nations, teachers lobbying for lower class sizes). The degree to which the voting public perceives the program benefits them is obviously a key issue for political support of targeted programs (Gelbach and Pritchett, 1997) but I want to hold that issue to one side for know.

[6]I assume the three groups have equivalent income so that marginal utility is equal.

TABLE I Model Structure of Public Groups.

		Public		
Verbal description	*Model parameter*	*Core supporters*	*Middle*	*Hard headed (economists)*
Prior belief in the efficacy of the action (gain to program beneficiaries (in own group utility) per dollar expenditure)	α	High	Middle	Low
Weighting of program gains versus own consumption	γ	High	Middle	Low
"Persuadability"	β	High	Middle	Zero

Note: Model structure: key groups of the public and their distinguishing characteristics.

In the absence of an evaluation their beliefs about program effectiveness are a combination of their initial beliefs and the amount spent by advocates on persuasion (S)[7]:

$$\text{If no evaluation, beliefs are: } \alpha_i = \alpha_i^0 + \beta_i * S \tag{5}$$

This simple set up characterizes the three groups of the public by (a) their weighting of program gains (γ), (b) their initial beliefs in program efficacy (α), and (c) how amenable their beliefs are to promotional activities of advocates – for which I use the neologism: "persuadability".[8]

2.3.1 Core Supporters

This groups weights program gains high relative to own consumption. There are two interpretations of this. The easiest to formalize is that this is high altruism. The alternative, more cynical, perhaps more realistic, but also harder to model, explanation is that core supporters are either those who consider themselves likely to be program beneficiaries or are the suppliers who will benefit from an expansion of the project. "Core supporters" have high initial beliefs about program efficacy and are easily swayed by persuasion.

2.3.2 Hard Headed

This group has a relatively low prior belief about program efficacy and is completely unpersuadable; their prior beliefs cannot be swayed by anything less than a randomized experiment. Finally, I also assume the hard headed are also hard hearted – so that program gains weigh less in their utility than the other groups.

2.3.3 Middle Group

In between the core supporters and the hard headed is the middle group, which, as the name suggests, is in between the two groups on each of the three parameters: they have modest

[7]Note that spending is a public good in the sense that the magnitude of spending to convince each group is independent of the size of the group. Alternatively persuasion spending could be per person, which would obviously change the role of group size in decision making.

[8]I adopt an ugly new word to avoid a loaded word to describe how amenable people's beliefs are to promotional/persuasive activities. If one calls it "gullibility" then this obviously takes one stance while referring to it as "receptiveness" conveys an entirely different viewpoint.

concern for program benefits, believe public interventions can be effective, and can be swayed by evidence short of a randomized experiment.

Since we assume (for simplicity) the vote is over an additional dollar from each person if we express the budget in terms of dollars per person in the population there are four possible budget outcomes. The outcomes depend on the proportion of each group in the population and whether for each group their beliefs about program efficacy exceed their threshold level.

2.4 When Will Politically Savvy Altruists Do Evaluations?

So, knowing what the voting public will do in the second period, what should the advocates do in the first? The advocate's beliefs about the findings of the level of program efficacy if a rigorous evaluation program efficacy could fall into the four ranges of the *critical values* for support from the various groups of the public from Table II:

$$
\begin{aligned}
p_A^0 &= P\{\alpha^{\mathrm{RE}} \leq \alpha_{\mathrm{CS}}^C\} \\
p_A^{\mathrm{Low}} &= P\{\alpha_{\mathrm{CS}}^C < \alpha^{\mathrm{RE}} \leq \alpha_M^C\} \\
p_A^{\mathrm{Med}} &= P\{\alpha_M^C < \alpha^{\mathrm{RE}} \leq \alpha_{\mathrm{HH}}^C\} \\
p_A^{\mathrm{High}} &= P\{\alpha_{\mathrm{HH}}^C < \alpha^{\mathrm{RE}}\}
\end{aligned}
$$

An advocate would favor a rigorous evaluation only if the advocate's expected utility from doing the evaluation was higher than welfare with the optimal level of spending and no evaluation:

$$
\begin{aligned}
\text{Do evaluation if: } & p_A^0 U_A(0 - C) + p_A^{\mathrm{low}} U_A(E^{\mathrm{Barebones}} - C) + p_A^{\mathrm{Med}} U(E^{\mathrm{Operational}} - C) \\
& + p_A^{\mathrm{High}} U_A(E^{\mathrm{Full}} - C) > U_A(E(S^*) - S^*)
\end{aligned} \tag{6}
$$

The simplest case is where the advocate knows for sure that the evaluation will reveal true program efficacy. Thus the question amounts to the following: if true efficacy is known, will the politically determined program expenditure level, net of evaluation costs, exceed the net program expenditures achievable through optimally chosen spending used to persuade the various groups of the public?

The crucial decision is how much to spend on promotional activities. For a given distribution of initial beliefs there is a level of promotional spending (possibly zero) for both the core supporters and the middle group that will just raise the group's actual belief in program efficacy to the critical level. Given the assumed simple linear assumptions about beliefs

TABLE II Outcomes Based on Program Efficacy Beliefs.

Proportion of the population supporting funding	*Description of funding level*	*Beliefs in program efficacy*
0	No program (or termination)	$\alpha_{\mathrm{CS}} < \alpha_{\mathrm{CS}}^C$
$F_{\text{core supporters}}$	Barebones	$\alpha_{\mathrm{CS}} > \alpha_{\mathrm{CS}}^C$, $\alpha_M < \alpha_M^C$
$F_{\text{core supporters}} + F_{\text{middle}}$	Operational	$\alpha_M > \alpha_{\mathrm{MS}}^C$, $\alpha_{\mathrm{HH}} < \alpha_{\mathrm{HH}}^C$
1	Full funding	$\alpha_{\mathrm{HH}} > \alpha_{\mathrm{HH}}^C$

Note: Program budget outcomes under different beliefs about program efficacy.
α_i^C is the critical level of beliefs about program efficacy as described in Eq. (2).

(Eq. 5) depends on the gap between initial and critical level and the persuadability for each group:

$$S_i^c = \frac{\alpha_i^c - \alpha_i^0}{\beta_i} \tag{7}$$

For the hard headed group $\beta = 0$ so there is never any promotional spending (only rigorous evaluations can persuade them), but for the other groups the optimal level of spending (conditional on spending) depends on whether the level of spending necessary to garner their support exceeds the incremental budget (which in this simple case with pure "public good" effect of promotional spending and one dollar of budget from each person is the share of the group in the population).

$$S_i^* = \begin{cases} 0 & \text{if } S_i^c > F_i \\ S_i^c & \text{if } S_i^c \leq F_i \end{cases} \tag{8}$$

Doing a rigorous evaluation has the drawback that it may lower the mean belief about efficacy – sufficiently to erode program support – relative to what could have been achieved by promotional activities. So the question is whether the benefits – essentially avoiding promotional costs – are worth the costs of an evaluation.

This depends on the allocation of groups in the population (which directly determines the differences in expenditures), the initial beliefs of the groups, and how amenable their beliefs are to promotion. Table III works through the possible configurations of true program efficacy, persuadability of the public, and cost of randomized evaluation.

The first two cases (I and II) contain sub-cases where politically savvy altruists may well be called *cynical altruists*: advocates claim program effectiveness higher than their beliefs (which is the true value) and would resist evaluations and instead prefer to "pilot and persuade". In case II (b) advocates believe the program is not sufficiently effective as to win support if the middle group knew the true value of program effectiveness. The reason for this is not necessarily that efficacy is "low" in some absolute sense, it could also be that the middle group altruism (relative to the objectives and beneficiaries of the program) is low. Savvy advocates can plausibly reason:

> "*If the middle group cared more, as they morally should, about the issue of education/nutrition/drug abuse/homelessness/health/AIDS and/or had more humane/ethical/lofty degrees of general altruism then the* ***true*** *program effectiveness would be sufficient for them to support the program (after all it is enough for us) – but they don't. However, fortunately the middle group is sufficiently persuadable that it is cheap enough to convince them that the program effectiveness is high. So the best approach is to never do an evaluation that would reveal the truth but rather maintain sufficient uncertainty about true program efficacy so that we can overstate program benefits sufficiently to garner middle group support.*"

For these advocates the issue with evaluations is not feasibility or cost. Over some ranges of parameters advocates will *not* do an evaluation *even if it were free* – they would be willing to pay to avoid a randomized evaluation.[9] In this case it truly "pays to be ignorant" – advocates have higher utility if no one knows the truth about program efficacy.

2.5 Comparative Statics: What Makes Evaluations More Likely?

This is a model that explains why, given the beliefs of single objective, single instrument advocates about the outcome of an evaluation and the initial beliefs and persuadability of various groups advocates would prefer not to have a rigorous evaluation of the impact of

[9]The key word being "complete" as there are many dollars taken to "do" evaluations.

TABLE III Decision to Evaluate.

Case	*Configuration of advocates (certain) beliefs about program impact and the various groups initial beliefs about program impact*	*Conditions on persuadability of the public (critical values of spending for each group to attain support) and cost of randomized evaluation)*	*Advocate chooses evaluation or "pilot and promote"*	*Resulting program expenditure level (Tab. II)*
I	$\alpha^{\text{True}} \leq \alpha^{C}_{\text{CS}}$ (No support from any group if true value is known)	(a) $S^{C}_{\text{CS}} > F_{\text{CS}}$	Pilot, zero promotion	Termination
		(b) $S^{C}_{\text{CS}} \leq F_{\text{CS}}$, $S^{C}_{M} > F_{M}$	Pilot and promotion	Barebones
		(c) $S^{C}_{M} \leq F_{M}$	Pilot and promotion	Operational
II	$\alpha^{C}_{\text{CS}} < \alpha^{0}_{\text{CS}} < \alpha^{\text{True}} \leq \alpha^{C}_{M}$ (Support from "core supporters" at true value)	(a) $S^{C}_{M} > F_{M}$ (Persuasion costs higher than middle group gain)	Pilot, zero promotion	Barebones
		(b) $S^{C}_{M} \leq F_{M}$ (Persuasion costs low)	Pilot and promotion	Operational
III	$\alpha^{0}_{MS} < \alpha^{C}_{M} < \alpha^{\text{True}} \leq \alpha^{C}_{\text{HH}}$ (Support from "middle group" at true value)	(a) $S^{C}_{M} > C$, $C < F_{M}$ (Persuasion costs high, evaluation cost lower than middle group gain)	Evaluation	Operational
		(b) $S^{C}_{M} \leq C$, $S^{C}_{M} < F_{M}$ (Persuasion costs lower than evaluation cost, middle group gain)	Pilot and promotion	Operational
IV	$\alpha^{C}_{\text{HH}} \leq \alpha^{\text{True}}$	(a) $S^{C}_{M} > C$, $C < F_{M} + F_{H}$ (Persuasion costs high, evaluation cost lower than gain)	Evaluation	Full funding
		(b) $C > S^{C}_{M} + F_{H}$, $S^{C}_{M} < F_{M}$ (Evaluation costs too high, persuasion costs for middle group less than gain)	Pilot and promotion	Operational

Note: Decision by single issue, single instrument advocates to perform rigorous evaluation depends on true value of program effectiveness, persuasion costs, and evaluation cost.

their proposed intervention. In many situations advocates would prefer ignorance to knowledge. I believe that these situations are common – because the gap between the altruism of advocates and the public can be large. It is hard to know how to marshal evidence, but I suspect that evaluations are rare because the middle group has low altruism and few interventions have sufficient efficacy to satisfy them relative to the level of efficacy required by advocates and core supporters (which may included providers and beneficiaries who benefit directly). In this case (essentially in the classification above – made more likely by uncertainty) there will be many programs operating and promoted and lobbying for middle group support – but resisting evaluation. But this is not a model that predicts there will never be evaluations. Let us now examine the conditions that make an evaluation more likely. Since in cases I and II there are no conditions in which programs are evaluated the first question is whether the "state of the (model) world" places us in case I and II or in II and IV.

2.5.1 Evaluation is Less Likely When Altruism is Small

In cases I and II advocates will *never* do a rigorous evaluation. The necessary condition for being in either of these cases is that true program efficacy is less than the critical value necessary for middle group support. For any given "true" value of program efficacy, case II is more likely than case III when the middle group's altruism is lower. That is, from Eq. (3) the middle group will support the program when the truth is known only when the marginal utility delivered to program beneficiaries is "large" relative to marginal utility of own consumption of the middle group, where "large" is determined by the degree of middle group altruism:

$$\text{Support if: } \alpha_{\text{True}} > \alpha_M^{\text{Critical}} = \left(\frac{\partial U}{\partial X}\right) * \left(\frac{1-\gamma_M}{\gamma_M}\right) \tag{9}$$

The derivative of the critical value of program efficacy for middle group support is:

$$\frac{\partial \alpha_M^{\text{Critical}}}{\partial \gamma_M} = \left(\frac{\partial U}{\partial X}\right) * \left(-\frac{1}{\gamma_M^2}\right) \leq 0 \tag{10}$$

So that decreases in the altruism coefficient of the middle group make it less likely that any given program really meets their criteria. Suppose for instance the "core supporters" were perfectly altruistic so that at $\gamma = 0.5$ and they would provide support if the program were to provide exactly one unit of marginal utility to the beneficiaries per dollar of expenditures. Now suppose that the middle group's utility was such that $\gamma = 0.1$ so that if marginal utility of the middle group and core supporters was equal then they would have to believe that program efficacy was nine time as high as that of the core supporters. To be even a value of 0.1 seems relatively high – suppose that altruism was quite small and $\gamma = 0.01$, then program efficacy would have to be 99 times its own marginal utility to justify program support.

If one believes that altruism is not in fact strong in the general public then this would suggest that empirical conditions will nearly always be in case II – outside of a group of core supporters program advocates have to rely on persuasion rather than evaluation to convince the middle group. If altruism is low very few programs will in fact be sufficiently effective to garner political support if the truth were known so the optimal strategy will nearly always be persuasion and not rigorous evaluation.

2.5.2 Evaluation When the Truth Helps the Advocates

While within case III or IV evaluation is a possibility, it is only one possibility, and what makes program evaluation more likely *within* case III will be examined. In case III true program efficacy attracts middle group support ($\alpha^0_{MS} < \alpha^C_M < \alpha^{True}$) but this only means that advocates *might* choose to do an evaluation. But evaluation is still done only if (a) it is cheaper to achieve the support to expand the program through an evaluation than through persuasion and (b) securing the additional budget is worth it. Still in the case where the core group needs no support in the certainty, linear utility model, the condition for an evaluation is that:

$$\text{Evaluate if } C < S_M^{\text{Critical}} = \frac{(\partial U/\partial X) * ((1-\gamma_M)/\gamma_M) - \alpha^0_M}{\beta_M} \quad \text{and} \quad C > E^{Op} - E^B = F_M \tag{11}$$

Figure 1 illustrates the basic comparative static exercises on the amount it takes to convince the middle group in the absence of a rigorous evaluation. A rigorous evaluation is more likely:

- *The more initially skeptical middle group.* The lower the initial belief in efficacy of the middle group (α^0_M) relative to the "truth" the more likely it will be in the interests of the advocates to carry out an evaluation rather than spend money to persuade (Fig. 1). One has to speculate on the psychology of the middle group to say how this affects any given program, but one can see where there are some remedies that are more "intuitive" than others. For instance, suppose the middle group shares the objective of reducing fertility – it is plausible that public provision (or subsidization) of contraception to achieve that objective has intuitive public appeal as contraception is an obvious *proximate* determinant

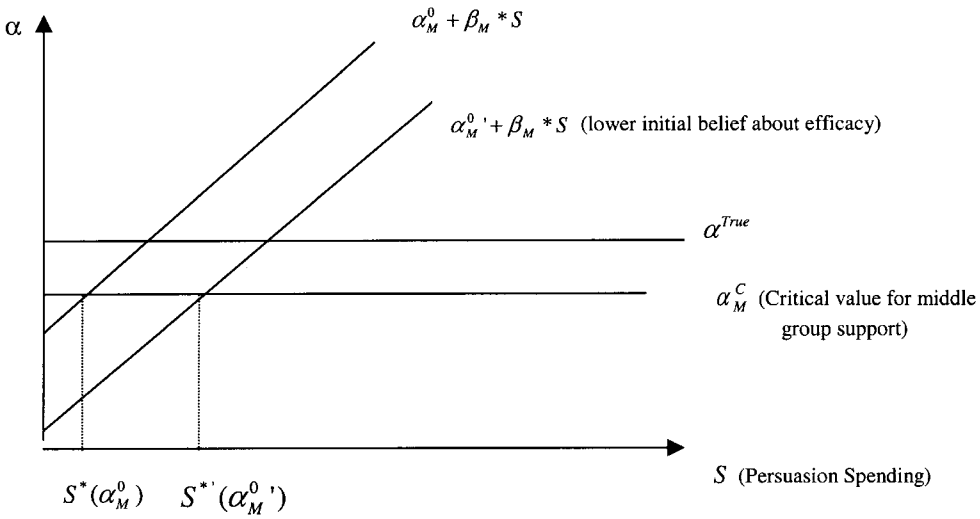

FIGURE 1 If the middle group is initially more skeptical (lower initial belief in program efficacy) the persuasion spending necessary for middle group support (S*) increases, making a rigorous evaluation of cost C more likely.

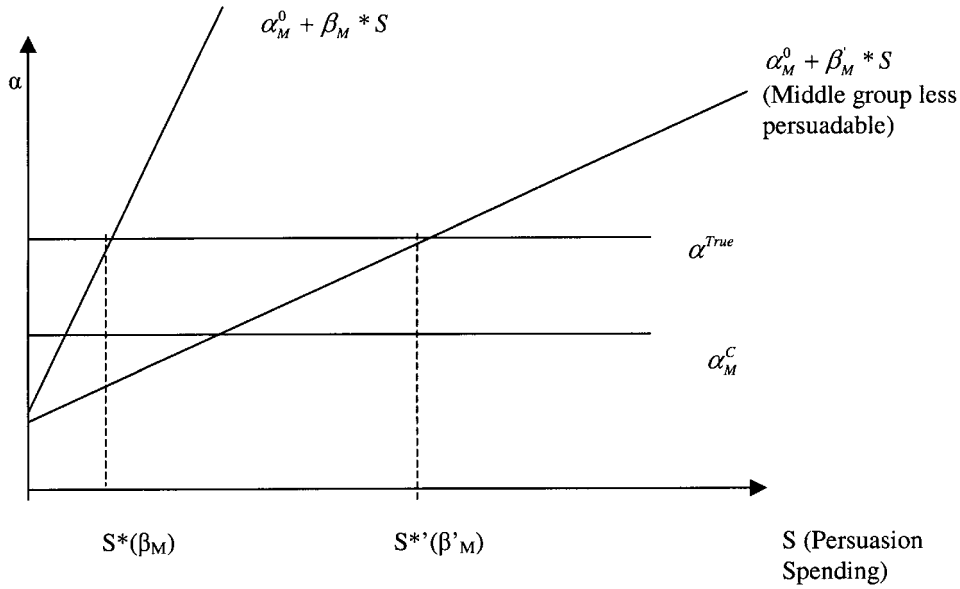

FIGURE 2 As the middle group is less persuadable the spending necessary to achieve support increases, making an evaluation more likely if the "true" value is sufficient to attract middle group support.

of fertility. Other actions for reducing population growth which achieve their impact through a more complex chain of causation (such as increasing female education) may have less initial credibility of program efficacy.

- *The larger the budget gain from convincing the middle group* (*the smaller the group of core supporters*). Even if evaluation is to be preferred to persuasion it still has to be the case the middle group is worth convincing. If the core group is very large (perhaps because of self-interest of program beneficiaries) then it may be the case that the advocates would do neither an evaluation nor promotional activities.

2.6 Nervous Altruists: When Advocates Are Uncertain About Evaluation Accuracy

Uncertainty about the outcome of the randomized evaluation will create another complication: *nervous altruists*. One reason to fear a randomized evaluation is that the evaluation may not – for methodological reasons, complexity, mis-measurement, or just bad luck – capture the true benefits of the program. In this case there is the risk of a "bad draw". In this case advocates may believe that the true program efficacy is actually sufficiently high that the middle (or even hard headed) group would support the program if the rigorous evaluation revealed the true value – but they may fear the evaluation would produce a lower value. With risk averse advocates the optimal decision might be to not do an evaluation even when they believe the program is effective enough even relative to the middle (or high) group's levels of altruism. In this case of the *nervous altruists* "it pays to be ignorant", even if when they were confident the evaluation would reveal the truth they would prefer a rigorous evaluation over "pilot and persuade".

It is not worth working out all of the cases with uncertainty over outcomes and risk aversion. Instead I will work out the intuition in one case. In particular how do outcomes change in case

III(a) (in which the promotion to achieve middle group support was sufficiently expensive that it was preferable to evaluate) with the introduction of uncertainty and risk aversion?

The simple intuition in this sub-case is that with risk aversion at the same level of evaluation cost and S^* at which advocates would have chosen evaluation, they will prefer the certain outcome under S^* to facing the uncertainty of an evaluation.[10] So, unless the upside risk is large (because either the probability of a "high" evaluation outcome is large or the "hard headed" population is large – $p_A^{\text{High}} * F_{\text{HH}}$) the advocate would prefer the certain outcome to the risk of an evaluation.

3 COALITIONS OF SINGLE OBJECTIVE ADVOCATES: A PLEDGE OF SECRECY?

But, alas, ignorance is not bliss. While some advocates are convinced both of their issue and their remedy, in many other instances advocates must form coalitions and exist in more complex institutional settings.

There are many instances in which advocates are certain about the *objective* they regard as important, but are not certain about the instrument to best addresses that objective. By not evaluating any programs advocates might gain funding but they themselves are uncertain about the impact. Single objective, multiple instrument advocates – such as those interested in population reduction, malnutrition, crime, etc. – want more funding but would also like more information about the relative merits of different instruments.

Suppose there are three advocates, two single objective, single instrument advocates named A and B (one who favors program "A" and one of whom favors program "B") and third advocate named C who is passionate about the objective but who is agnostic between instruments A and B. Under what conditions will these advocates form a coalition and belong to the same organization and what would such a coalition do?

One outcome is an organization that has a pledge to secrecy. Advocates A, B and C form a coalition with the agreement that they will do a rigorous evaluation of the instruments A and B and then raise monies and fund the most effective program (all of this assumes the "core group" support can be brought into the coalition). All of this will be done with a pledge of secrecy so that only the advocates themselves are allowed to know the actual, cardinal, program efficacy. To the outside world only the *rankings* of the programs are revealed. So in its money raising efforts the organization claims: "we are only funding the most effective programs." If the middle group's persuadability is higher when the advocacy comes from a multiple instrument organization then there are conditions in which organizational coalitions will form around issues because this allows them to raise more funds than they would otherwise. This essentially allows them to move from case II(a) in which only a "barebones" level of funding is possible to a case in which middle group support and hence the "operational" level of funding is available for at least one of the programs as illustrated in Figure 3.

The key to organizational stability is that no one cheats on the secrecy agreement. But there are two obvious ways to cheat on the agreement. First, is to promise a rigorous evaluation but not complete it until after your coalition partner had completed theirs. At that point, if the rigorous evaluation estimate of program efficacy is lower than that necessary for program support the "cheating" partner can say: "all we can credibly promote without

[10]Of course this assumes that the only uncertainty is over the outcome of the evaluation and that all other parameters are known with certainty by the advocates. But of course if advocates are also uncertain about the middle group's persuadibility then nothing can be said about the impact of uncertainty in general.

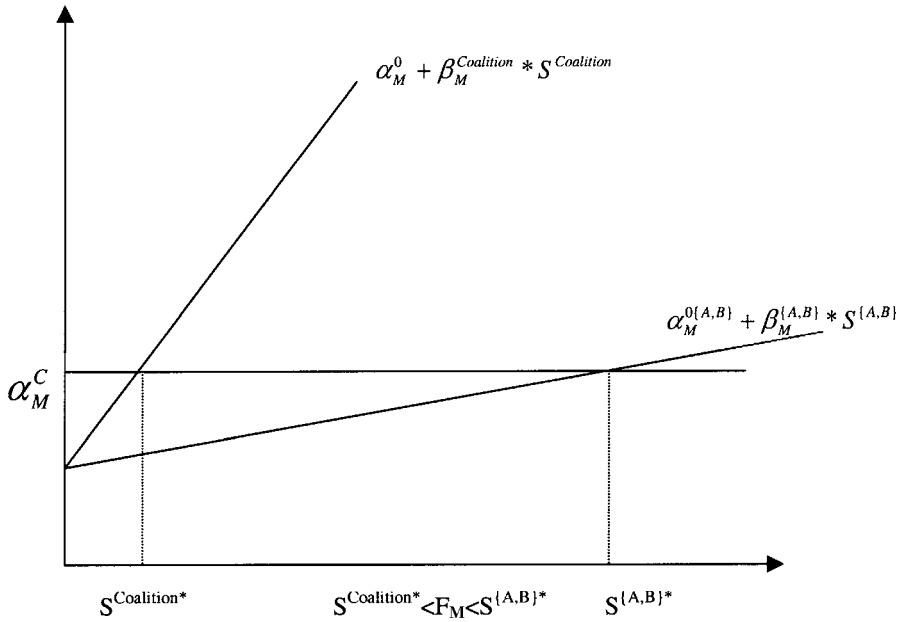

FIGURE 3 Case in which a coalition sufficiently increased middle group persuadability so that the better of intervention A or B is funded even though neither of them singly would receive middle group support.

complete cynicism and without fear of exposure is my preferred instrument (about which we are still ignorant) and we can only do that as long as I don't complete my evaluation so let's stop now and promote my intervention." Alternatively, suppose the evaluations reveal that program B is more effective and hence there is no funding for instrument A. But if advocate A leaks the *absolute* level of program efficacy for program B this creates a situation such that the organization can no longer raise middle group support for program B but it can still raise support for program A (as long as advocate B does not retaliate and leak the results of the evaluation of program A). Given this *ex post* cheating on the original agreement (to evaluate and fund only the most effective) by advocate A it would then be in the best interests of all three (A, B and C) to continue and advocate for program A. Of course this would make for a very unstable coalition because (i) *ex ante* B and C might suspect A's sincerity, (ii) if the organization requires repeated interactions then B and C are unlikely to be fooled again, (iii) advocate B might be tempted to behave "irrationally" and quit the organization and retaliate by leaking the results of the evaluation of program A – leaving them all worse off.

4 MULTIPLE ISSUES, MULTIPLE INSTRUMENTS: UNSTABLE COALITIONS, NAÏVE TECHNOCRATS

It is a waste of money to add an evaluation component to nutritional programs – these evaluations never find an impact anyway – we should just move ahead with what we nutrionitsts know is right.

Nutritional advocate in project decision meeting.

It's amazing how many bad projects get support. Epistemologically, why do you think that is?

email from a colleague

While the objectives of some organizations are clear and the only question is efficacy in pursuing those objectives, others have a broad objective that covers multiple issues (and within each there are multiple possible instruments). A development bank (World Bank, ADB) or aid agency (USAID, DFID) are examples in which the organizational mission ("development" or "poverty reduction") is sufficiently broad that a variety of *issues* might contribute (basic education, infrastructure, health, trade liberalization, judicial reform). Now not only is there a question of which instrument is best to pursue a given sub-objective, but there is also the question of how any given sub-objective contributes to the overall objective – which is "more important" health or transport?

So, for instance, take the organization I happen to know best, the World Bank. Suppose the World Bank's mission really is poverty reduction but that internally, among the staff, and externally, among the governments who are the organizations shareholders, there is a very loose coalition of general and specialized advocates that supports the mandate of "poverty reduction".

Within the organization there are advocates of all kinds: some single objective, single instrument, some single objective but with no fixed beliefs about the efficacy of specific instruments, some have no strong beliefs either about the relative importance for poverty reduction of sub-objectives and or about the efficacy of instruments within sub-objectives. Table IV tries to capture just a small fraction of this complexity – as there are layers upon layers of what might constitute "sub-objectives" and what with the sub-objectives groups might have strong feelings about as instruments. The categories listed in the table are obviously not mutually exclusive. In addition there are "thematic" concerns like "gender" which I cannot illustrate as they cut across all individual issues/sub-objectives and instruments.

All the stakeholders in the coalition have their own agendas and specific ideas about the most effective sectors and/or instruments. I assume that for many of the advocates with the organization they would not stay within the organization (working on a different

TABLE IV Coalitions with Broad Objective.

Z: Poverty Reduction most important objective

A: Education most important
- **B:** *Class room construction*
- **C:** *Teacher training*

D: Economic growth most important
- **E:** *Trade liberalization*
- **F:** *Infrastructure investment*

G: Agricultural productivity most important
- **H:** *Irrigation*
- **I:** *Land reform*
- **J:** *Extension*

K: Health most important
- **L:** *Child vaccinations*
- **M: AIDS**
  - **N: AIDS**
    - **P:** *Prevention high risk groups*
    - **Q:** *Prevention mass education*
  - **O:** *Treatment*

Note: Each letter represents a "stakeholder" with given views about importance for poverty reduction of various "sub-objectives" (in **bold**) and the strength of their belief about particular instruments for sub-objectives (in *italics*).

An example of coalitions of advocates within an organization with a broad objective – the World Bank and poverty reduction.

issue/instrument) if it allocated zero resources to their preferred issue/instrument. I refer to this state as a loose coalition. So if the World Bank devoted resources to education but zero resources to teacher training person C would leave but not A or B. If the World Bank were to devote zero resources to AIDS persons M, N, O, P, Q would leave but person P would leave even if resources were devoted to Health, AIDS prevention but zero were allocated to prevention in high risk groups.[11] This "loose coalition" nature of the organization implies that a "pledge of secrecy" is not possible.

In this set up, only the most naïve would voluntarily undertake a rigorous evaluation of their preferred objective/instrument. An evaluation would restrict the claims the advocate could make. This makes them a sitting duck for those advocates within the organization who have not done an evaluation and who would like more of the institutions resources for their preferred activity. Any evaluation will reveal weaknesses and a lower impact than could have been plausibly claimed for some other instrument. For instance, doing a rigorous evaluation of social funds sets them up to be attacked by supporters of specific sectors (education, roads, micro-credit) and vice versa.

So if there is to be any rigorous evaluation it would have to come from agent Z in Table IV – the stakeholder with poverty reduction as the objective who is agnostic about sub-objectives and instruments – rather than emerging spontaneously. If agent Z is "management" they would need evaluations in order to optimize the use of resources. Assume that organizational resources were fixed and there was a consensus on how to measure the objective of poverty reduction and if the magnitude of resources were fixed then the optimum would be reached if resources were allocated to the activity with the highest poverty reduction gain, which consists of two parts: how effective program/policy "A" is in producing gains in the sub-objective per dollar of input, and how much the sub-objective contributes to the overall objective of poverty reduction. So assuming a set of mutually exclusive actions A in set $\{A\}$ of equivalent cost and funds for only one of them the decision rule should be:

Decision rule for organization with poverty reduction as an objective (with secure funding):

$$\max_{\{A\}} \left(\frac{\partial \text{ Poverty Reduction}}{\partial \text{ Sub-objective}} \right) * \alpha^{A}$$

But this simple rule ignores the *trade-offs*. There are three elements to the trade-off faced by the management of a broad mandate organization: raising the average efficacy of actions undertaken so that resources actually do contribute to the general objective, maintaining persuadability of the general public so that the organization is more effective in resource mobilization than the individual advocates alone would be, and sustaining the internal coalition behind the broad objective.

The only reason for the individual advocates to remain in a coalition is if they can better pursue their objectives inside than outside the organization. One reason for this is that the organization is better able to mobilize resources, in part because it is easier to mobilize "middle group" support for a coalition because persudability is higher. This suggests doing sufficient evaluation activity to maintain organizational credibility.

Doing some evaluation maintains appearances of effectiveness, but doing compelling and convincing evaluations of an intervention strongly favored by an important constituency group runs a huge risk: what if relative to other interventions it is not in fact effective? But doing a rigorous evaluation without a pledge of secrecy implies that activities will be cut,

[11]This is not a statement about relative magnitudes of support as the division into issues and instruments are arbitrary and for illustration only.

which will undermine support for the broad objective as the advocates and "core group" supporters for those particular activities are lost to the broader organization.

In order to know either element of the optimal fixed budget decision rule with any degree of quantitative precision and certainty one would have to do a rigorous evaluation. Without a rigorous evaluation the average efficacy of the funded activities is lower than the optimal.

The solution to this general problem is going to be specific to the organization and will be a complex mix, but almost certainly will involve a great deal of strategic ignorance. The "real world" solution is likely to be one in which the organization tries to generate just enough evaluation to allow the experts and "general issue" advocates to push the organization towards the more effective interventions and maintain persuadability of key stakeholders, but not so much evaluation that any sector/intervention that needs support for political reasons would be ruled out. Moreover, there would definitely not be enough evaluation to assess rigorously the overall level of efficacy of the broad organization.

5 VERISIMILITUDE

The model I am proposing here explains the dearth of reliable, rigorous, evaluations of a variety of public and non-profit actions as a strategic commitment to ignorance by the advocates of these programs.[12] Ironically, I cannot think of any way to rigorously test my claims about the causes of the lack of rigorous tests of claims. Why, in spite of that, do I think this model helps understand reality? Both less cynical and more cynical explanations fail and because there is a certain amount of verisimilitude to the model.

5.1 Less and More Cynical Explanations

Perhaps there are other reasons for the lack of rigorous evaluations than the strategic behavior by advocates: cost, ethical barriers, practical impossibility have been proposed as alternative explanations.

That evaluation is "too expensive" is often cited as an explanation for its scarcity, but since evaluation costs are a tiny fraction of program costs and the potential gains are enormous it is difficult to believe this is a compelling reason for substantial areas of public intervention.[13]

Ethical issues seems a poor explanation of why there *are* evaluations in medicine where it is a matter of life and death and *are not* evaluations of, for instance, educational innovations. Because of budget constraints very few people have access to the intervention (*e.g.* small class sizes, free family planning) or the program/project in any case so the randomization does not deny anyone "access" to the program – budget constraints do. Finally, one would think the ethical issue of "policy malpractice" (Stigler, 1974) through perpetuation of ineffective action is at least as serious an issue as structuring participation in programs of unknown efficacy in order to learn if they are effective.

That randomized evaluation is "not feasible" is usually simply false. Moreover, even in instances in which it would be easy to structure a randomized evaluation (for instance, where

[12]The same is true of their private sector counterparts and the existence of persuasive (as opposed to informative) advertising expenditures in the absence of rigorous information suggests strategic ignorance is not limited to the private sector. I have yet to see a rigorous evaluation of the impact of breath mints.

[13]In the report on the RAND Health insurance experiment (Manning *et al.*, 1987) the authors point out that in spite of the large cost of the experiment the savings from implementation of even one of the policy recommendations stemming from the report would pay for itself in about a week!

program implementation is going to be phased in due to logistical constraints and the sequence of areas receiving the intervention could be randomized) these measures are resisted.

Then there are *more* cynical explanations of ignorance – which is that pure political interests are at play. A more cynical explanation of the lack of rigorous evaluation of educational innovations could be built purely around the self-interest of educators than that there are altruists who honestly believe in the efficacy of their intervention. While there is obviously some cooperation between advocates and their "core groups" – so that teachers' unions will fund advocacy for class size reduction – I do not believe that all advocacy is pure self-interest, most advocates are quite sincere in their beliefs.

5.2 Reasons To Believe

There are several interesting facts about the evaluations that do occur. Although I obviously do not have a complete sample of all, the evaluations suggest that something like the present model is of strategic interest.

First, a huge number of evaluations are started and very few are finished, written up, and publicized. This evaluation attrition is too large to be consistently "bad planning" and is more likely strategic behavior.

Second, I do not have a complete sample, but many programs that have had randomized evaluations were in fact eliminated – and it is not clear whether this had anything to do with the evaluation or not. The voucher program in Colombia was eliminated before the randomized evaluation results were even available. The training program evaluated under JTPA was terminated. The fact they were eliminated *ex post* at least suggests these program were without solid political support and the evaluation itself was a strategy of weakness.

Third, randomized evaluations are often implemented by those out of the mainstream, groups with much less to lose if the outcome is adverse. For instance, the randomized evaluation of the provision of textbooks in Kenya was carried out by a small NGO – not the government (Kremer *et al*). The implementation and evaluation of the Colombia voucher program was not carried out in the Ministry of Education (King *et al*).

Fourth, it is interesting to look at the pressures behind the evaluations that do exist, and typically one finds that the proposed intervention was either not supported by the "core supporters" or had strong opposition otherwise.

6 CONCLUSION

Who *really* wants to know? While serendipity plays some role in knowledge, most increases in knowledge about the impact of public programs or policies are the result of deliberate research. If a program can already generate sufficient support to be adequately funded then knowledge is a danger. No advocate would want to engage in research that potentially undermines support for his/her program. Endless, but less than compelling, controversy is preferred to knowing for sure the answer is "no."

This strategic ignorance of advocates is not necessarily a bad thing. Just as Frank Knight argued that economies will have a higher growth rate when its entrepreneurs are more irrationally optimistic, it is possible that a combination of excessive subjective certainty among altruistic advocates and strategic maintained ignorance of true program effects is actually welfare improving. Pritchett (2001) shows in a Rawlsian model in which people can make binding decisions from behind a pre-birth position in which no one knows what their status will be, welfare is higher if all actors promise not to do rigorous evaluations. Perhaps

all of us quantitatively oriented public policy social scientists are violating a pre-birth commitment to our fellow human beings.

Acknowledgements

I would like to thank various people, who are not ignorant, for comments: Alberto Ades, Julie van Domelen, Jeffrey Kling, Susan Stout, Jeffrey Hammer, Deon Filmer, and Scott Guggenheim.

References

Angrist, J., Bettinger, E., Bloom, E., King, E. and Kremer, M., Vouchers for private schooling in Colombia: Evidence from a randomized natural experiment, NBER Working Paper 8343.

Barnett, W. S. (1996) Lives in the balance: Age-27 benefit-cost analysis of the High/Scope Perry Preschool Program (Monographs of the High/Scope Educational Research Foundation, 11). Ypsilanti: High/Scope Press.

Filmer, D., Hammer, J. and Pritchett, L. (2000) Weak links in the chain: A diagnosis of health policy in poor countries, *World Bank Research Observer*, **15**(2), 199–224.

Glewwe, P., Moulin, S. and Kremer, M. (1997) Textbooks and test scores: Evidence from a prospective evaluation in Kenya. Mimeo: Harvard University.

Manning, N., Duan K., Leibowitz and Marquis (1987) Health insurance and the demand for medical care: Evidence from a randomized experiment, *American Economic Review*, **77**(June), 251–277.

Phillips, J. F., Simmons, R., Koenig, M. and Chakrabarty, J. (1988) Determinants of reproductive change in a traditional society: Evidence from Matlab, Bangladesh, *Studies in Family Planning*, **19**(6), 313–334.

Schweinhart, L. J., Barnes, H. V. and Weikart, D. P. (1993) Significant benefits: The high/scope perry preschool study through age 27 (Monographs of the High/Scope Educational Research Foundation, 10). Ypsilanti: High/Scope Press.