8

Forecasting the Effects of Scaling Up Social Programs: An Economics Perspective

Robert A. Moffitt

The problem of scale-up, or forecasting the effects of interventions at a larger scale than that for which their estimated effects were originally obtained, occurs across many different applications, programs, and disciplines. Economic models of scale-up, which are the concern of this chapter, have focused on particular types of scale-up effects that occur frequently in interventions where economic outcomes are the major interest and where program beneficiaries—in this case, students, rather than teachers or schools—are usually the actors making the decision of whether to take up the intervention. However, because they have the ambition to provide a general model of individual choice behavior, economic models have a much wider applicability than to economic outcomes alone and to the decisions of students alone. Economists have indeed begun in recent years to apply their models to broader sets of outcomes and issues, such as the effect of educational interventions on noncognitive outcomes, the importance of peer effects within schools, and the effects of special education programs. However, this work is still relatively immature as a subfield and in addition, economists have done almost no work, and have developed almost no models, for certain types of scale-up effects, particularly those concerning the change in the nature of the intervention itself, the effect on which program operators most often focus.

It is argued here that economics has nevertheless much to contribute to the problem of scale-up. First, the economic model of production processes provides a natural framework within which to discuss the problem of scale-up in general, and to develop a taxonomy of different types of scale-up effects. As an example of its usefulness, I argue below that it allows one to provide alternative explanations for one of the most common findings in

the scale-up literature, namely, that effects at larger scales always seem to be weaker than at smaller scales. Second, the economic model provides particularly good insights into some, but not all, of the types of effects listed in such a taxonomy, particularly those having to do with scale-up effects in inputs and outputs. Third, the economic model has led to a general framework for empirical evaluation research and causal inference that can be usefully employed in the measurement of scale-up effects, particularly by nonexperimental means using natural variation. While none of the fundamental problems of measuring scale-up effects are "solved" by the econometric models, these models do provide suggestions for an approach and a framework within which evidence can be accumulated and progress can be made.

This chapter is not concerned with the question of how a researcher can get a successful small-scale program to be adopted by a larger set of schools, how it can be managed at a larger scale, why some interventions appear to "spread" and others do not, or what characteristics of a successful small-scale intervention are mostly likely to result in its being taken to scale. As important as these questions are, they require an analysis of how schools and institutions actually adopt innovations, and this is beyond the scope of this chapter and indeed, they are not questions (perhaps unfortunately) that economists in the evaluation literature have generally considered.[1] This chapter is instead concerned with the scientific question of how to forecast the actual effects of an intervention prior to its being adopted at a larger scale.

The chapter first lays out a conceptual, economic model within which scale-up effects can be discussed, and then provides some discussion of those effects on which the economic model has something to say. Then issues of measuring and estimating scale-up effects (i.e., the forecasting problem) are discussed.

SCALE-UP CONCEPTS

As stressed by Hedges in his chapter, conceptual models are necessary to make progress on the problem of scale-up. Purely statistical models alone are unlikely to be satisfactory because there are too many causal effects involved in the scale-up problem, and purely statistical models will most likely not adequately separate the different confounding factors and individual effects that are at work. Like all difficult problems where the complexity of the real world is much greater than the data and the methods at our disposal, having a theoretical framework to guide thinking and to interpret the data concerning the scale-up problem is essential.

Production Function Model

The production function model is very familiar to education researchers and needs no elaboration, for it has been used repeatedly as a framework within which the effects of educational inputs on student outcomes can be understood (Levin 2001; Lazear 1999). It has its critics as a useful model to understand the nature of the educational process (Hanushek 1986; Mayston 1996), but here it will be used in a more general way to describe the nature of the mechanism by which individuals are drawn into treatments and later enter a posttreatment state, with selection mechanisms at work at both ends. Figure 8.1 illustrates the simplest such model. A population of individuals exists, from whom a subset are drawn into the program and receive the intervention. It is probably sufficient to define the population as the "eligible" population although this can be deceptive if the criteria for eligibility are endogenous, for in that case the size and nature of the eligible population can change as the program is scaled up. There is a process, defined as a specific set of treatments applied to a set of individuals (possibly differentially by individual characteristics), which constitutes the intervention. Individuals emerge at the other end and outcomes are observed for them individually, and the distribution of outcomes for the entire exiting group is observed as well. Those who drop out of the intervention prior to its completion are included in the exiting group, and their outcomes are regarded as part of the outputs of the intervention, even though their effects may be zero or close to it. Outcomes are subdivided into short-run and long-run outcomes; this by itself is an innocuous distinction but is useful because scale-up effects differ along those dimensions, as discussed below.

The paradigmatic case is that in which estimates have been obtained on a small program, but interest centers on its effects when the program is put in place in a larger area, such as city- or statewide, or even nationally. For example, a curricular innovation has been tested and found favorable in the schools in one area but now is being considered for adoption statewide. Typically one obtains from the small-scale evaluation (whether experimental or

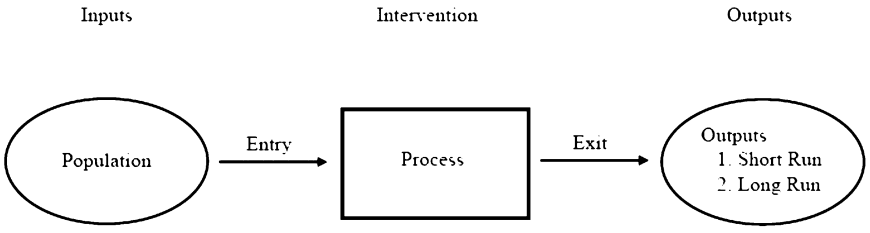

Figure 8.1. Production Function Model

nonexperimental) estimates of the effect of an intervention on some outcome variable *Y* of some population *P*. The effects may differ for those with different values of a set of individual characteristics, or contextual factors, *X*. The population *P* represents characteristics of the sample in the evaluation in addition to *X*, and often is measured by some indicators of the nature of the process by which individuals enrolled in the program and how they were selected to be in it. Often participants are joining the program on their own accord, but they may also be referred or required to participate. Statistically, we can say that the small-scale evaluation provides estimates of the function $E(Y|T,X,P)$ for different values of *T* and *X*, but generally for only a single value of *P* (though this may vary as well, as described below).

Economists distinguish the concept of scale-up from the larger problem of generalizability, or external validity, but educational researchers in this area often do not do so. For example, moving from a small scale to a larger scale may result in an enrolled population with a different set of individual characteristics or in areas with different contextual factors (*X*) from that of the individuals and areas used for small-scale estimation. Perhaps the educational innovation was tested on children in a middle-income school but it is being considered for adoption in a low-income school. Or the innovation was tested on students of largely one ethnic or racial group and is being considered for adoption in schools where a different ethnic or racial group constitutes the majority. Economists term this a problem of generalizability because the effects of the innovation may differ for students of different income levels, or different racial and ethnic groups, and therefore an extrapolation problem must be solved when forecasting the effects of the innovation to the different group or different context (assuming no direct estimates are available for the different group). Economists do not consider this to be a problem of scale because it exists even in cases where scale is not an issue—that is, where the innovation's effects were measured in a small area (e.g., one school) and the innovation is being considered for adoption in a different small area (e.g., a different school with different students and context) and one is trying to forecast the effect in the different school. Economists instead reserve the term *scale-up* for problems of generalization that have a change in scale as an intrinsic element, and that would occur even if the types of individuals or schools involved in the initial evaluation were the same as those in the areas where adoption is being considered. This chapter concentrates on problems where scale-up is an intrinsic issue, and does not use as examples problems that involve attempts to forecast the effects of innovations on different areas or types of individuals per se. However, in practice both problems tend to occur together, for large scale-up almost always involves bringing under the intervention some types of areas or individuals who were not in the tested areas, as well as changes of scale itself. The empirical problem of forecasting is, however,

the same problem, and this will be discussed below when that problem is addressed.

Scale-Up Effects in Inputs

Table 8.1 lists a taxonomy of scale-up effects that will be discussed here. The table divides the effects into those pertaining to inputs, those pertaining to the intervention, and those pertaining to outputs, and distinguishes between short-run and long-run effects.

In the category of inputs, a short-run scale-up effect occurs if there is some voluntary element to participation in the program and if knowledge of the program diffuses through the population rather than occurring instantly. Such effects can occur whether the program in question is completely new, and its impact was estimated initially on only a small set of individuals drawn into the program by some special process, or it is an existing program where a reform has been made and it is the effect of the reform on entry that is the issue at hand. Some reforms can conceivably be viewed unfavorably by many in the population, in which case diffusion of information about it may reduce entry rather than increase it.

Thinking about how individuals in the population will view the new program or the reform raises immediately the important question of how individuals, or schools in some cases, would come to be enrolled or involved in the program after scale-up. The long-run effect in table 8.1 of change in entry mix reflects the fact that the individuals or schools involved in the large-scale program may differ in some way from those in the estimation sample in ways that could not be measured in the latter. One example that often comes up in purely voluntary programs, where individuals or schools make their own participation decisions, is that the estimation sample is often conducted on individuals or schools that are particularly advantaged or

Table 8.1. Taxonomy of Scale-Up Effects

	Effects	
Stage of Production	*Short Run*	*Long Run*
Inputs	Knowledge of diffusion	Change in entry mix Migration Other endogenous responses
Intervention	Knowledge build-up on best technology for intervention	Change in nature of the treatment Change in resources per recipient
Outputs	Lags in effects of output scale-ups	Market responses Social interactions Policy-institutional reactions

disadvantaged relative to the population as a whole. For example, interventions are often initially conducted on a particularly disadvantaged sample. An intervention that is aimed at very disadvantaged individuals but which, after scale-up, brings into the program less-disadvantaged individuals who benefit less from the program, will result in a dilution of the program effects when measured as an average. This is one effect that is consistent with the commonly observed reduction in intervention effects when going to scale (mentioned in the introduction), and can occur even if the intervention, or treatment itself, is unchanged after scale-up; merely the composition of the enrolled population may change.

If the selection mechanism involves some voluntary elements, then another possibility that arises is that the individuals or schools considering participation may be able to obtain information on the effectiveness of the program, and make their decisions on that basis. If they do, and if they perceive, rightly or wrongly, that the effectiveness of the program is different after scale-up from before, that too can affect entry and the composition of the enrolled population after scale-up. This will be mentioned again below in the discussion of output effects.

The nature of these effects will differ depending on selection. If those administering the program do not allow purely voluntary participation, then the question is how the selection mechanism will change after scale-up. That question has to be answered on a case-by-case basis, depending on the application in question.

Economists have been relatively successful in constructing plausible, and empirically verified, models of the voluntary participation decisions of individuals in social programs. The standard model for such effects is some kind of benefit-cost calculation, either expected utility maximization or some related concept. Economists have made less progress in modeling the decisions of program operators in deciding whom to admit to a program in those cases where enrollment is not entirely voluntary. Modeling the decision process of organizations is much more difficult than modeling that for individuals.

The relevance of entry and diffusion effects to classroom innovations of various kinds is still present but is likely to operate in a different fashion because individual students cannot select themselves in and out of a classroom where an intervention has been implemented. An exception occurs when parents, guardians, and/or teachers lobby effectively on behalf of individual students to include or remove them from classrooms selected for implementation of the intervention during scale-up. Nevertheless, there could be effects of classroom innovations on the nature of the students entering that classroom if those innovations affect curriculum or student or teacher behavior at earlier grades, or if schools likewise make alterations. A small-scale intervention in only one classroom may not affect school or

teacher policy in earlier or later grades, but a large-scale intervention may. It is one of the paradoxes of entry effects that entry mix effects are likely to be small if the intervention itself is small and incremental, and does not have a large impact on outcomes; but the more successful the intervention, and the larger the effects on outcomes, the more entry mix is likely to be a problem.

It is also difficult to separate measurement from theory in this case because whether these types of spillover effects into earlier classrooms are a problem depends on whether the "small-scale" intervention took as its unit of observation the student in the classroom where the intervention was taking place, or the school. To the extent that the school was the unit of observation, and the innovation was implemented "schoolwide," the effects mentioned above may very well be captured.

Table 8.1 also lists migration and other endogenous responses as long-run input effects. If migration occurs as individuals move into (or out of) the areas where the intervention is offered, or across areas because the intervention differs across those areas, this can also generate a scale-up effect that is not captured by the small-scale estimates. Individuals moving into a school district where a particularly successful intervention has been brought up to scale—or out of a district where the intervention is of a type that some parents dislike—may, likewise, change the input mix and therefore the average effectiveness of the program in question. This is really a subcategory of the entry mix problem. Other endogenous responses of this kind are possible, such as changes in personal or family characteristics to make oneself eligible for a program (income, family structure, etc.).

Scale-Up Effects in Intervention

Many practitioners think of scale-up effects as occurring primarily in the nature of the intervention itself. This effect is most often described as the problem of "implementation," meaning getting the program operators (in this case, teachers and schools) to actually implement the program in the same way it was implemented in the small-scale test. In some of the discussions in the educational scale-up literature, where this problem is considered to be overwhelmingly the most important one, a "successful" educational innovation is defined not only as one that has a positive effect on outcomes of students in the small-scale test, but also as one that is easily implementable by schools and teachers in the larger educational system.

The economics literature on program evaluation discusses several reasons for the presumed importance of implementation. One is the general notion that it is more difficult organizationally to administer a program to a large group of individuals than to a small group; this is another explanation for a diminution of effects when going to scale. However, this notion needs to

be parsed and some important distinctions, though perhaps only conceptual ones, need to be made. Administering a program to a large number of individuals does not technically require any different treatment process than administering a program to a small set of individuals provided the technology of the treatment is kept the same; for example, if the intervention is administered to groups of the same size as in the small-scale program (meaning necessarily more groups). There may be administrative difficulties higher up in the organization that may yield inefficiencies, but this is a very different type of effect. In many cases, instead, the notion that the program is harder to administer to a large set of individuals than to a small set arises because the technology is not held fixed and the treatment given at the individual level in the small-scale program is not replicated at the larger level. Sometimes this can be thought of purely as a resource issue because a smaller amount of resources per enrollee may be devoted at the larger scale than at the smaller scale. That individuals might be treated more uniformly, and with less personal attention, in a large-scale rather than a small-scale program is an example of the treatment's actually changing when going to scale.

These effects are listed as long-run effects in table 8.1, and it is fair to say that economists have not studied these issues much, partly because they are so difficult. To do so properly requires a model of how treatments are administered at different scales, and how the nature of an intervention changes with scale, and this is an inherently difficult problem. Many small-scale evaluations do conduct "process" evaluations or studies of how services are actually delivered in a particular small-scale intervention (the Bloom, Hill, and Riccio chapter in this volume is one of the better illustrations of this type of work), but one of the weaknesses of most analyses of that kind is that they are not fed into any type of structured model that could be used for extrapolation and generalization to other, and larger-scale, types of programs.

A rather different, short-run effect of scale-up in the intervention occurs when those running the program change the nature of the treatment (in a positive direction, presumably) as more effective ways of serving the population are continually discovered. Programs are rarely static and unchanging, and new programs in particular almost always evolve over time. Nevertheless, this is listed as a short-run problem in table 8.1 on the presumption that the program will eventually stabilize if left in place long enough, and it is this long-run effect that is of most interest to the evaluator.

Scale-Up Effects in Outputs

Economists have conducted the majority of their work on scale-up effects in outputs, which are sometimes lumped together as "general equilibrium"

effects. The textbook example is that of a market response that occurs when an intervention becomes large enough in scale to affect supply and demand in a market and hence, changes the equilibrium price. In many examples, the price response to a large-scale intervention acts to reduce the average effect of that intervention—for example, increases in the supply of more skilled labor reduces its equilibrium wage—making the estimate from the small-scale intervention too large. This provides a third possible explanation for the commonly observed diminution in program effectiveness after going to scale (in addition to the dilution of the nature of the entry pool, and reduction in the effectiveness of the treatment itself for a constant entry pool). Economists are well equipped to study clearing of markets and to consider the multiple feedback effects that can occur when an intervention is large enough to affect markets.

The relevance of these effects to classroom interventions below the adolescent years is questionable, because the main type of effect studied by economists is the effect of the productivity of the school-leaving pool on the youth labor market. Interventions that were large-scale and close to the school-leaving point, and which had a large enough effect to (for example) increase the skill level of graduates could conceivably have an effect of this kind. However, the more important "general equilibrium" effect of this kind for classroom innovations is its effect downstream, that is, on classrooms at later grades. A truly successful intervention that improves the cognitive skills of students in a particular dimension will undoubtedly have effects on how material is taught in later grades, and this will not be captured by a tested intervention that is so small in magnitude as to not affect the average skill level of students in the upper grades.

But there are two other scale-up effects in outputs, aside from the classic market example, which are potentially important as well. One is the presence of social interactions, as they may be called, which arise only when a program is scaled up. An example is the development and establishment of social norms and expectations that arise when large numbers of individuals undertake a treatment and become aware that others have done so. The feedback effects so generated make the small-scale impact estimates invalid. If the feedback effects are positive in sign, this is one case where the small-scale intervention may underestimate the large-scale effect because the latter reinforces and extends the former by propagation through the larger student population. Another example is where the individuals affected by an intervention affect the outcomes of individuals not in the program. For example, students whose performance improves may have positive effects on the performance of students who have never been in the program if they are in the same classrooms or have some other type of social contact. Peer effects are one specific example of such effects. Economists have recently begun to model these kinds of effects but have made only modest progress to

date (Kremer and Levy 2003; Nechyba 1996; Epple, Figlio, and Romano 2004).

Another even larger-scale output scale-up effect occurs if the institutional or policy environment changes in response to the scale-up of the program. Typically this is of concern only when the intervention in question is a very large-scale, structural change in an entire program or system. Examples include welfare reform in the United States in the mid-1990s, and possibly the No Child Left Behind legislation. The effects in question here arise if programs other than those that have been affected change their service offerings in response to the reform of the initial program. In the case of welfare reform, if new child-care programs spring up, if the nature of job training programs changes to serve a different clientele, if new tutoring or remedial programs are created after the intervention, or other changes in the local policy environment occur, these truly "macro" effects can also affect individual outcomes and therefore cause the small-scale estimates to be invalidated.

All output effects can have effects on inputs if the effectiveness of, or payoff to, the program affects program entry decisions. Programs that have some voluntary element, for example, can be expected to bring in more enrollees if the program is perceived as successful as it is unsuccessful. Likewise, changes in the treatment discussed earlier can affect entry decisions if the nature of the treatment is correctly perceived by the population and there are voluntary elements to enrollment.

MEASURING SCALE-UP EFFECTS

Measuring scale-up effects is a difficult task and requires departing from the standard experimental or nonexperimental model, both of which consider the impact of a treatment on a set of individuals or organizations holding constant the scale of the program, the entry pool, the nature and implementation of the intervention, and the scale of the output effects. Therefore measurement must go in other directions.

Because evaluation methodology becomes important in the discussion of measuring scale-up, the following discussion separately considers experimental, natural variation, and simulation methods. In all cases, it is assumed that valid small-scale estimates of the effect of an intervention on outcomes for a particular population are available.

Experimental Methods

The typical small-scale randomized field trial (RFT) does not capture scale-up effects. Generally, the typical modification in experimental method-

ology to capture scale-up effects is to conduct experiments at the community level and to make them saturation experiments. Thus, randomizing a set of areas or school districts into treatment and control groups would almost by definition capture most entry scale-up effects (except for immigration from other areas), and at least some output scale-up effects (although not those market responses that occur in other areas), and will include some intervention-related scale-up effects. In short, by testing a program by implementing it on the entire population of an area, it is possible to obtain a direct estimate of the total effect of a program, thus capturing scale-up. If a large number of areas is tested, this approach essentially is a partial implementation of the actual program on a large scale, and consequently it is not surprising that it should capture many scale-up effects.

Unfortunately, there are many difficulties with implementing this idea and, as a result, it is rarely a viable option. One problem is that enrolling a sufficient number of areas to gain a reasonable level of statistical significance is extraordinarily costly and beyond virtually all research budgets. The common practice of pairing single comparison areas with single treatment areas is subject to too much variability to be reliable, and there are many examples in areas of social welfare intervention where comparison-site designs have proved faulty because of random events in one of the two areas. A second difficulty that often arises is simply a political one, for it is often difficult to obtain the cooperation of large numbers of political entities in a randomized trial, at least in our decentralized government where mandates from the top are rarely possible. A third difficulty is that controlling the treatment to make it homogeneous across the areas is always quite problematic. For all these reasons, a statistically reliable saturation-side experiment to capture scale-up effects is a nonstarter.

Natural Variation

Some types of input and output effects can be captured by statistical analysis using natural variation in scale across areas. Using this variation, however, does require the construction of some type of statistical model that can relate the scale-up effects to the effects available from the small-scale estimation. The small-scale estimation will provide, for example, "first stage" estimates of the effect of the intervention on outcomes of the individuals enrolled. A statistical model is then required to relate the effects of such a change in outcomes, generalized to a larger population, that work through feedback, either through market responses or social interactions. Estimating those feedback effects is possible with nonexperimental data, using natural variation across areas in other dimensions. Market output responses, for example, require estimates of the price responsiveness to a shift in a supply or demand curve, and there is an extensive econometric literature on how to

estimate such types of relationships with natural variation using observational data. Social interaction effects, while much more difficult to measure, can in principle be measured with the right kind of exogenous, cross-area or cross-group variation in the mix of individuals with different outcomes, allowing estimation of peer effects and social norm effects.

Entry effects are more difficult to measure because small-scale estimation typically provides no information at all on how individuals would come to be enrolled in a scaled-up program. Exceptions sometimes occur when small-scale estimates are available for different areas, or for different sets of individuals; these exceptions at least provide some estimates on how the outcomes will differ for a different input mix (the population P referred to previously). A model of entry is required, and one must develop a model of how individuals or organizations make choices to participate in similar programs which can be extrapolated or mapped into the entry effects of the program in question. The "similar" natural variation may be difficult to locate in existing programs or past evaluations, but this is required to capture input mix effects.

There are definite limits to these types of exercises, however, for they work only in some circumstances and with the availability of natural variation in the first place. They are typically not possible for scale-up effects that occur in the intervention, where there is rarely direct natural variation or relevant natural variation in a related treatment which can be used instead, but only in output and input effects. Imprecision in many nonexperimental estimators of this kind, and the threats to internal validity which arise so frequently, further weaken this approach.

Simulation

In many cases the only, or most promising, approach, is to construct a theory-based simulation model that can be used to forecast the magnitude of the scale-up effects of various kinds. Entry mix, market and social interaction responses, and even how the nature of treatments varies with scale can in principle be formally modeled. Calibrating such simulation models is the difficult part, and must rely on previous estimates obtained from natural variation to inform the values of the parameters assumed. In some cases, there may be no reliable estimates of parameters in the simulation model, in which case the best that can be done is to simulate with a plausible range of parameters and to leave the final estimates uncertain and only falling into a range. Theory-based simulation is also only as good as the theory used to construct the models, and some theories have been validated more than others from past research. Nevertheless, there are many ways to quasi validate simulation models from outside data to ensure that they are correctly representing at least existing, or historical, behavior, and this al-

lows such models to be grounded more firmly than they would be otherwise.

Different Populations and Contexts

Finally, as noted previously, the problem of generalizing the estimates of a small-scale intervention to areas or schools with different types of students or different contextual factors is not a scale-up problem per se but can nevertheless be likewise discussed under experimental, natural variation, and simulation methods. Experimental methods would seem to be very appropriate here—and feasible, provided that sufficient numbers of schools can be persuaded to test an innovation. The recommendation for multisite designs made by Hedges in his chapter in this volume is exactly aimed at obtaining information on how the effects of an innovation that was successful in one particular school in one particular area would differ for different schools, students, and areas. Multisite designs would have the additional advantage of providing information on the much-discussed problem of what educational innovations are "adaptable," meaning that they can be implemented successfully in different schools and in different populations than those in the initial study. Natural variation is the nonexperimental counterpart to randomization, where natural variation in some type of school innovation is necessary for estimation. Simulation is the ultimate solution if nothing else is available; in this case, a model of how treatment effects differ by student, school, and area would have to be developed, using as a research base the knowledge gleaned from past studies of other educational interventions on how impacts vary along those dimensions. Once again, the uncertainty inherent in this type of forecasting would require sensitivity testing and the production of a range of estimates rather than a single one.

SUMMARY

Analyzing scale-up effects is difficult and requires different models and methods than those used for the typical small-scale evaluation. In light of the difficulties involved, it is important to begin by conceptualizing the problem correctly, forming a taxonomy of different scale-up effects and relating them coherently to one another. Indeed, constructing a theoretical model of scale-up is important simply to organize any empirical approach to the problem. None of the empirical means of measuring scale-up effects is particularly attractive, but the approach most likely to yield insights—though not "solutions"—is a simulation model based on theory, which is informed by the collection of empirical estimates available or which can be

reliably obtained from nonexperimental analysis, possibly from using natural variation across areas.

NOTE

1. The area of economic research where a somewhat related set of issues has been discussed is the literature on incentives in organizations and, to some extent, game theory. Very little of this literature has specifically focused on organizational determinants of the adoption of innovations, however.