Series Foreword

The Tjalling C. Koopmans Memorial Lectures were initiated by the Cowles Foundation in 1989 using a fund established by Koopmans' family, friends, and colleagues shortly after his death in 1985. These lectures offer an opportunity for preeminent scholars of economics to provide synthesis and perspective on a body of ongoing research to a broad audience of economists. The Cowles Foundation is pleased to introduce a monograph series based on these lectures, with the first volume written by Ken Wolpin, the Walter H. and Leonore C. Annenberg Professor in the Social Sciences at the University of Pennsylvania.

Tjalling Koopmans was a prominent voice in the early efforts of the Cowles Commission (later Cowles Foundation) to bring rigorous logical, mathematical, and statistical methods of analysis to the study of economics. His pathbreaking work on linear programming methods and their application to problems of optimal resource allocation led to his Nobel Prize in economics (jointly with Leonid Kantorovich) in 1975. He also made important contributions to the foundations of empirical economics, including his work on identification of structural economic relationships. Beyond his own research, Koopmans' collaborations and leadership of the Cowles Commission/Foundation influenced the paths of many important scholars in economics.

Especially relevant to the topics of the present volume was Koopmans' 1947 article "Measurement without Theory." Although written as a book review, it offered a critical assessment of an entire research program that aimed to build up knowledge through documentation of empirical regularities, unencumbered by economic theory. Koopmans' view arose not from a preference for theory per se, but from recognition of the inherent limitations of empirical analysis that is neither guided nor restricted by an economic model. He viewed theory as indispensable for revealing what quantities should be measured, how they can be measured, and how measurements can be interpreted.

In this volume, based on lectures given at Yale in November 2010, Ken Wolpin offers a modern perspective on the role of theory in empirical economics. Ken's work has followed Koopmans' vision closely, although on a range of topics Koopmans could hardly have anticipated in 1947. Ken has made major contributions to labor economics, economic demography, development economics, health economics, and empirical methodology.

Here he begins by considering what is perhaps the most common motivation for going beyond mere description of the data: *ex ante* policy evaluation. Using prominent examples from the literature, Ken illustrates the essential role theory plays in exposing which variation in the data can reveal the policy effects of interest. This is combined with a pragmatic view of the unavoidable trade-offs between the assumptions one makes in empirical work and what can be learned from the data given these assumptions. In the second part of the monograph, Ken moves to a sequence of extended examples, each drawing on a significant applied literature. These examples allow him to illustrate in detail the types of guidance empirical economists gain from theory, both in providing (or challenging) the logical foundation of an empirical strategy and in revealing the proper interpretation of results.

Ken's lectures stimulated a great deal of discussion among my colleagues. I thank him for the substantial work involved in transforming a pair of talks into a monograph that could be shared more broadly.

Philip A. Haile
Cowles Foundation Director, 2005–2011

The Koopmans Memorial Lectures

Menahem Yaari, 1989
Janos Kornai, 1991
Thomas Schelling, 1993
Alain Monfort, 1994
Peter A. Diamond, 1996
Paul R. Milgrom, 2004
James Heckman, 2006
Lars Hanson, 2008
Kenneth Wolpin, 2010

Acknowledgments

I am grateful to the many colleagues who have helped to shape my thinking about empirical methodology. In this regard, I am particularly fortunate in having had long-term collaborations with Zvi Eckstein, Michael Keane, Mark Rosenzweig, and Petra Todd. I have also benefited from interactions with the numerous graduate students with whom I have worked. Finally, I would like to thank the Cowles Foundation for giving me the opportunity to present these lectures.

1 Introduction

The two lectures that comprised my Tjalling C. Koopmans Memorial Lectures were titled "*Ex Ante* Policy Evaluation" and "The Limits of Inference without Theory." I have chosen the second as the title of this book because it encompasses a broad theme that is also illustrated by the content of the first lecture. The title obviously borrows from the famous Koopmans (1947) essay "Measurement without Theory" published in 1947. That essay was a direct response to the book by Burns and Mitchell (1946), *Measuring Business Cycles*, and, more generally, to the extensive data collection effort being conducted at the National Bureau of Economic Research on business cycle fluctuations. Koopmans supported the enterprise of collecting data to better understand business cycle fluctuations but argued that the productivity of that enterprise would be significantly enhanced if it were guided by theory. In his concluding remarks, he stated: "But, the decision not to use theories of man's economic behavior, even hypothetically, *limits* the value to economic science and to the maker of policies, of the results obtained or obtainable by the methods developed."[1]

The purpose of these lectures was not to reopen the debate about the proper use of theory in developing facts (see, for example, Christ, 1994; Kydland and Prescott, 1990) but to consider instead the role of theory in drawing inference from data and the limits that eschewing the use of theory places on inference. My intention is to illustrate the applicability of Koopman's concluding remark to inferential empirical work in economics and in the social sciences more generally. The discussion focuses on microlevel research, that is, where heterogeneity at the individual level is explicitly recognized. It is important to be clear about what empirical work is to be characterized as inferential. By that, I mean generally any research that uses data to draw conclusions that go beyond the mere statement of fact, that is, the tabulation of statistics.

There are (at least) two views about why theory and inference need not be connected. In one view, theory is seen as unnecessary or even detrimental to inferential data analysis. This view is exemplified in the expression that researchers should "let the data speak for themselves." This approach is often called "reduced form," although that terminology is a distortion of its original meaning, which stemmed from the work by Koopmans and other members of the Cowles Commission. The second view is more nuanced and arguably more sound. In this view, in a first stage of analysis, inference can be made without the use of theory. However, the ability to draw inferences without the use of theory requires what is called a "quasi-experimentalist" approach that involves implicit random assignment or an actual experiment, a randomized controlled trial (RCT).[2] Theory's role is then to provide an explanation for the inference that is derived from the experiment and as an aid toward generalization.[3] The absence of theory in inferential empirical work is pervasive. For example, of all the papers in the January 2009 maiden issue of the new American Economic Association journal *Applied Economics*, all of which were inferential, none contained an explicit model of "man's economic behavior."[4]

The reduced form and experimentalist approaches are often contrasted to what is popularly called the structural approach. In keeping with that nomenclature, I will refer to reduced form and experimental approaches as nonstructural. The structural approach is one in which the relationships that are estimated are explicitly intended to be invariant to policy, where policy is interpreted broadly to include counterfactual experiments. Under this definition, whether the relationship that is estimated is structural cannot be decided outside of a policy (or counterfactual) context. A variant of the structural approach is characterized as estimating "deep" parameters. In economics, deep parameters are usually associated with preference functions, technology, and constraints. In the first characterization, the parameters of interest may be combinations of deep parameters; deep parameters are by definition invariant to policy. The structural estimation approach requires that a researcher explicitly specify a model of economic behavior, that is, a theory. As Marschak (1953) describes it: "In economics, the conditions that constitute a structure are (1) a set of relations describing human behavior and institutions as well as technological laws and involving, in general, nonobservable random disturbances and nonobservable random errors in measurement; (2) the joint probability distribution of these random quantities."[5]

There is also a third approach to empirical work in economics that, although sometimes referred to as reduced form, is better termed quasi-structural. In that approach a formal model satisfying Marschak's first criterion is postulated, and in some cases comparative static (dynamic) results are derived. The relationships that are estimated are viewed as approximations to those that are, or could be, derived from the theory (for example, as based on the comparative statics). The parameters are functions of the underlying deep (policy-invariant) structural parameters in an unspecified way. The stochastic elements are also unspecified combinations of the "random quantities" in Marschak's second criterion. In contrast to this approach, the "reduced form" approach that eschews formal theory starts from a relationship that is to be estimated rather than derived. To be clear, what I mean by theory is not necessarily a mathematical model but rather a coherently expressed connection between assumptions and inference. Mathematics facilitates an assessment of coherency and also provides a connection between a model and its statistical specification, but mathematics is per se not necessary (nor, obviously, sufficient).

The structural/nonstructural taxonomy is not, in my view, the critical distinction in terms of empirical methodology. As was the theme of these lectures, the critical distinction is not the estimation approach but whether or not inference is theory based. Experimental work, whether based on an RCT, a natural experiment, or some other estimation approach, can in some instances be interpreted as seeking to recover a structural relationship, that is, one that is invariant to policy. However, the lack of an explicit theoretical framework, a common feature in nonstructural work, leaves the validity of any such interpretation to the reader.

In terms of organization, I begin with the first lecture on *ex ante* policy evaluation, highlighting the role of theory in that enterprise. I restrict attention to the case where there is no direct policy variation from which to extrapolate to new policies. Two structural approaches to *ex ante* evaluation are presented, nonparametric and parametric. In both approaches, the Marschak criteria are satisfied. The advantage of the nonparametric approach is that assumptions auxiliary to the behavioral model are unnecessary. The disadvantage of that approach is that it can be applied only for a restricted set of behavioral models. The parametric approach allows for richer models and thus for the consideration of a wider set of counterfactual policies but requires auxiliary functional form and distributional assumptions. The two approaches

are illustrated with two examples, a wage tax and a school attendance subsidy. The results from applications are summarized.

The second lecture presents a number of examples illustrating the limits of inference without theory. The first two examples in the chapter illustrate the importance of theory in econometric specification within the context of a (seemingly) quasi-structural estimation approach. They show more generally how the quasi-structural approach, lacking the discipline of the structural approach necessitated by having to specify the exact mapping between the theory and estimation and often appealing only casually to received theory, can lead to inferential ambiguity. The first example is relevant to dozens of papers spanning over 30 years of empirical research on unemployment duration, and the second to the even larger literature on the impact of public welfare on labor and demographic outcomes of women. The third example, that of estimating the effect of school attainment on earnings, illustrates the connection between theory and the recent econometrics literature on instrumental variables (IV) estimation with heterogeneous response. The motivating example in this case is based on the natural experiment approach to estimation. The fourth example illustrates the importance of theory in addressing questions of external validity in the context of a prominent field experiment in education. Each of these illustrations is placed within the context of the broader literature, and the examples are contrasted to recent contributions to their literatures that rely on theory for inference.

2 *Ex Ante* Policy Evaluation—The Role of Theory

The goal of policy evaluation falls into two categories: *ex post* evaluation and *ex ante* evaluation. The goal of *ex post* policy evaluation is to determine the impact of policies that have been implemented. This type of evaluation is ubiquitous in economics and in the social sciences more generally. The goal of *ex ante* policy evaluation is to determine the impact of prospective or "new" policies. New policies can extend existing policies either in the dimension of a particular policy parameter beyond its current domain or introduce new parameters to an existing policy. An example of the former would be an increase in the existing minimum wage and of the latter the introduction of a time limit on public welfare eligibility or the incorporation of a drug benefit into Medicare. A new policy may alternatively be entirely outside of the historical experience, for example, the initial introduction of the income tax or the provision of subsidies for school attendance as is now prevalent in many developing countries. *Ex ante* evaluation allows the policy maker to compare the impact of alternative policies prior to choosing one of the policies to implement, for example, raising the minimum wage by 50 cents, $1, or $10, imposing a 2-year or 5-year time limit on welfare eligibility, or choosing a particular proportional or progressive income tax schedule. Effective policy decision making is clearly enhanced by the ability to perform credible *ex ante* policy evaluation.

Evaluations of existing policies make use of actual policy variation. Evaluations of new policies that are extensions of existing policy parameters, for example, increasing the minimum wage, also make use of actual policy variation, although they require an extrapolation outside of the historical experience. Evaluations of new policies that cannot be based on actual policy variation, for example, introducing a time limit on welfare eligibility or a new program that subsidizes

school attendance, raise additional challenges and form the central focus of this chapter. Variation useful for *ex ante* evaluation must, in some way, provide an analogue to policy variation; that is, it must be policy relevant. The role of theory is to identify such policy-relevant variations.

The evaluation literature distinguishes between two types of evaluation methodologies: experimental and nonexperimental. The experimental methodology is based on an evaluation of a demonstration or pilot project in which participants and controls are randomly chosen. The nonexperimental methodology analyzes observational data using a combination of behavioral and statistical modeling. Each of these methodologies has been applied to *ex post* and *ex ante* evaluation.

The aim of the experimental methodology applied to a prospective policy is to determine whether and to what extent the goals of the policy would be met if implemented. In this case, the choice of the policy's characteristics, the "treatment," is determined by the experimenter. The experimental approach has been held to be the gold standard for evaluating policy. However, there has been considerable debate about that proposition (see, for example, Heckman and Smith, 1995; Deaton, 2009; Imbens, 2009). The debate has centered not so much on problems that can arise in practical implementation, for example, the effectiveness of the randomization, selective attrition, and Hawthorne effects, but rather on the conceptual issue of what it is that can and can not be learned from a social experiment. The advantage of a social experiment is that it identifies a policy effect for the population in the experiment, which, in the best of circumstances, has high internal validity. The disadvantage is that the experiment identifies the policy effect only for the treatment that is chosen and may not be applicable to other populations.

The aims of the nonexperimental methodology can be broader. The role of theory, which is often eschewed in the social experiments literature, is paramount. Theory identifies mechanisms that generate policy effects, and theory-based estimation quantifies the importance of alternative mechanisms that enable extrapolation to other policies and possibly to other populations. This payoff is achieved at the cost of untestable assumptions that are not necessary with social experimentation; in that sense, the nonexperimental methodology has lower internal validity. Just as practitioners of social experimentation must take care in designing and implementing their experiments, practitioners of theory-based estimation must provide evidence on model validity.

Natural "natural experiments," a term adopted by Rosenzweig and Wolpin (2000) to distinguish random events that arise in nature from other so-called "natural" or quasi-experiments, differ from social experiments in that the interpretation of the "policy" effect is not assumption-free. As is the case with instrumental variable estimates more generally, the implicit assumption in the natural and quasi-experimental literature is that (conditional on observables) the effect of the policy (or "treatment") is the same for all people (Heckman, 1997; Heckman, Urzua, and Vytlacil, 2006). The role of theory is to elucidate the assumptions that underlie interpretations of the policy effect.[1]

The development of methodological approaches to *ex post* policy evaluation using nonexperimental methods has been and remains an active area of research (see the recent survey by Todd, 2008). There is little methodological or applied research explicitly concerned with *ex ante* policy evaluation using nonexperimental methods, which is perhaps surprising given the potential benefits.[2] The purpose of this chapter is to summarize and provide examples of structural approaches to *ex ante* evaluation.

2.1 Structural Approaches to *Ex Ante* Evaluation

In this section I discuss two structural approaches to *ex ante* policy evaluation: nonparametric and parametric. For concreteness, the ideas are illustrated with two examples: (1) the effect of a wage (or earnings) tax on labor supply and (2) the effect of a subsidy conditioned on a child's school attendance. In both cases theory identifies surrogate variation that substitutes for the lack of explicit policy variation. In the first instance the wage itself provides the policy-relevant variation, whereas in the second, it is a combination of the child's wage and parental income. The discussion first centers on a nonparametric method for solving the evaluation problem. I then turn to a parametric approach that can accommodate *ex ante* evaluation under assumptions that the nonparametric approach cannot and that provides a potentially richer set of policies amenable to *ex ante* evaluation.

In the wage tax example I begin with the simplest economic model in which an individual is deciding on hours of work in a static context. A nonparametric matching estimator of the impact of a newly imposed tax is developed and illustrated for the case in which wages are observed for both participants and nonparticipants. The case in which wages are observed only for participants is shown to require a distributional

assumption on wage offers, although, as in the case of full observability of wage offers, it is not necessary to make an assumption about the functional form of the utility function. Several extensions of the labor supply model are considered, including introducing a fixed cost of work, child care costs (including an *ex ante* evaluation of a child care subsidy), and nonlinear taxes.

The second example, the introduction of a school attendance subsidy, begins with the consideration of a static model in which parents are deciding on whether to have their only child attend school or work in the labor market. It is shown that joint variation in the child wage and parental income can be used to obtain a matching estimator of the effect of the subsidy on school attendance without making functional form or distributional assumptions (with full observability of the child wage). It is shown how multiple children, when fertility is not a choice, can be accommodated but that the nonparametric estimator is no longer valid when fertility is a choice. The nonparametric matching estimator is shown to extend to a perfect-foresight life cycle model and, in some cases, to a dynamic imperfect-foresight model. The nonparametric matching estimator does not survive the introduction of child home production. In fact, the same matching estimator now corresponds to a different policy, one in which households are provided a subsidy if the child does not work in the market but either attends school or engages in home production. There is thus, in this example, an inherent ambiguity in the interpretation of the matching estimator that is model dependent.

The school attendance subsidy also serves to illustrate *ex ante* evaluation using a parametric approach. The analogue of the matching estimator of the subsidy effect is derived given a particular parametric specification of the utility function and a distributional assumption for the unobservable taste shifter. It is shown that with these parametric assumptions, *ex ante* evaluation can be conducted in the presence of home production, which was not possible nonparametrically. The static model is formally extended to the dynamic case, which fits into the framework of discrete choice dynamic programming (DCDP) models. A brief digression discusses the estimation methodology of DCDP models as a prelude to the presentation of parametric empirical applications of a school attendance subsidy program implemented in Mexico. Two DCDP models that differ significantly in their structures are presented and contrasted. Both use data from the Mexican PROGRESA experiment that provided a school attendance subsidy to a randomly selected set of villages.

Estimates of an out-of-sample prediction of the impact of doubling the PROGRESA subsidy are presented based on two nonparametric matching estimators and on the two DCDP models. The estimates of the DCDP models and one of the matching estimators are seen to be quite close. An interesting difference in the two DCDP approaches is that one of them uses only the control group for estimation and validates the model by predicting the subsidy effect obtained from the experiment. The other DCDP model uses both the control and treatment groups in the estimation. This difference in empirical approaches raises the question of whether the value of withholding a part of the sample from the estimation for the purpose of validation is worth the loss of estimation precision. This question is briefly explored in the final section of the chapter.

2.1.1 Structural Nonparametric Approach[3]

2.1.1.1 Wage Tax

Suppose that a policy maker is considering the introduction of a proportional wage tax and would like to know how labor supply and, thus, tax revenues will vary with the tax rate. As a place to start, assume that labor supply behavior can be described by a standard static model in which individuals choose the number of hours to work given their wage rate, their level of nonlabor income, and their available total time. The utility function is a twice-differentiable, quasi-concave function of consumption, c, and hours of leisure, $l = 1 - h$, where h is hours of work and total available time is normalized to one. The marginal rate of substitution between leisure and consumption depends on a set of observable preference shifters, X (a vector), and unobservable preference shifters, ε. Individuals have nonlabor income y and receive wage offers w. In the absence of a wage tax, each individual thus chooses hours of work to maximize

$$U(c, 1-h; X, \varepsilon) \tag{2.1}$$

subject to the budget constraint

$$c = wh + y, \tag{2.2}$$

where consumption is the sum of labor earnings, the wage times hours of work, and nonlabor income. Optimal hours of work can be derived as a function of w, y, X, and ε, namely

$$h^* = \phi(w, y; X, \varepsilon). \tag{2.3}$$

The hours function admits to both zero (nonparticipation) and positive hours. Zero hours occurs when the marginal rate of substitution between leisure and consumption exceeds the wage evaluated at full-time leisure ($h = 0$). Otherwise, the individual works positive hours. No additional assumptions about the form of the utility function are made.[4]

Introducing a proportional tax on wage income at rate τ $(0 < \tau < 1)$ alters the budget constraint to[5]

$$c = (1-\tau)wh + y. \tag{2.4}$$

The model with the tax is, however, isomorphic to the model without the tax in the specific sense that if $h^{**} = \eta(w, y, \tau, X, \varepsilon)$ denotes the solution for optimal hours with the tax, then

$$h^{**} = \eta(w, y, \tau; X, \varepsilon) = \phi(\tilde{w}, y; X, \varepsilon) \tag{2.5}$$

where $\tilde{w} = (1-\tau)w$.[6] Thus, the hours of work function without the tax (ϕ) is also the relevant function in the presence of the tax.[7] The implication of this observation is that the effect of introducing a tax τ on hours of work can be studied from *ex ante* population variation in wage offers.

To see that, note that the difference between the population mean hours worked (inclusive of 0 hours) for persons with wage offer $\tilde{w} = (1-\tau)w$, nonlabor income y, and observables X and persons with wage offer w and the same values of y and X is

$$E_\varepsilon(\phi \mid \tilde{w}, y, X) - E_\varepsilon(\phi \mid w, y, X) = \int \phi(\tilde{w}, y; X, \varepsilon) f(\varepsilon \mid \tilde{w}, y, X) d\varepsilon - \int \phi(w, y; X, \varepsilon) f(\varepsilon \mid w, y, X) d\varepsilon. \tag{2.6}$$

Under the assumption that, conditional on y and X, the distribution of unobserved preferences does not depend on the wage, that is,

$$f(\varepsilon \mid w, y, X) = f(\varepsilon \mid y, X),$$

the difference given in equation 2.6 is exactly the effect of introducing the wage tax on expected hours worked by a person with wage offer w, nonlabor income y, and observable X. In general, the effect of the tax on expected hours will differ depending on a person's w, y, and X. Thus, persons with the same wage offer but different nonlabor income and/or observable preference shifters will respond differently to the tax in terms of their hours of work, as will individuals with the same nonlabor income but different wage offers and/or observable preference shifters.[8]

Integrating over the joint population distribution of w, y, X, that is,

$$\int [E_\varepsilon(\phi \mid \tilde{w} = (1-\tau)w, y, X) - E_\varepsilon(\phi \mid w, y, X)] dG(w, y, X) \tag{2.7}$$

gives the effect of the tax on expected hours of work over the entire population. As seen from equation 2.7, different populations, those with a different joint distribution of ε, w, y, X, will exhibit a different mean hours of work response to the introduction of the tax.

The same analysis would govern the case of changing an existing tax, τ_1, which has not varied in the past, to a new value τ_2. In that case, the initial net wage would be $\tilde{w}_1 = (1-\tau_1)w$, and the new net wage $\tilde{w}_2 = (1-\tau_2)w$. The change in hours worked would be given by modifying equation 2.6 to $E_\varepsilon(\phi \mid \tilde{w}_2, y, X) - E_\varepsilon(\phi \mid \tilde{w}_1, y, X)$. Also, although the focus has been on hours of work as the main outcome of interest, the outcome of interest could also be the work decision, which is just a transformation of hours of work [i.e., $1(h^* > 0)$] or mean hours of work conditional on participation.

It is possible to obtain sample analogues of the population mean hours necessary for calculating the policy effect. Suppose data are available for a random sample of individuals on hours worked, nonlabor income, and the observables in X and also that wage offers are available for both nonparticipants as well as participants. The case in which wages are observed only for participants is considered below. There are several estimation approaches that can be taken.

One method would be to estimate the conditional mean hours of work function $E_\varepsilon(\phi \mid w, y, X)$ nonparametrically using a method such as kernel, local linear regression, or series estimation. Given an estimate of $E_\varepsilon(\phi \mid w, y, X)$, it is straightforward to calculate equation 2.7 given G. As with any nonparametric approach, extrapolation outside of the sample variation is not possible. Thus, one cannot compute $E_\varepsilon(\phi \mid \tilde{w}, y, X)$ if $\tilde{w}$ is not observed in the sample for some (or all) values of y and X. Given that responses to the tax will generally be heterogeneous, the sample estimate of equation 2.7 may deviate from the population effect.

Another approach is to use a matching estimator.[9] The key insight is the observation that a person with a wage w subject to a tax τ would, on average, choose the same hours of work as an otherwise identical person, in terms of y and X, with a wage of $\tilde{w} = (1-\tau)w$ who was not subject to a tax. Thus, in the absence of a tax, the difference in mean hours of those with wage $\tilde{w}$ and those with wage w (given y and X) provides an estimate of the tax effect on hours for those with wage w.

Taking the average over all w, y, and X provides an estimate of the population mean hours effect of the tax. An advantage of the matching estimator for recovering the policy impact relative to the nonparametric estimation of the conditional mean function is computational. The conditional mean function is more informative, but some of what is recovered is unnecessary for the *ex ante* policy evaluation.[10] The matching estimator, on the other hand, recovers only the policy effect, that is, the appropriate finite change in the conditional mean function due to the policy.[11]

The matching estimator of the policy impact of the tax on mean hours worked takes the form,

$$\hat{\Delta} = \frac{1}{n} \sum_{\substack{j=1 \\ j,i \in S_P}}^{n} \hat{E}(h_i \mid w_i = (1-\tau)w_j, y_i = y_j, X_i = X_j) - h_j(w_j, y_j, X_j), \tag{2.8}$$

where the matches can be performed only for the n individuals whose w values and associated $(1 - \tau)w$ values both lie in the overlapping support region, $S_P = \{\tilde{w} \text{ such that } f_w(\tilde{w}) > 0\}$.[12] The estimator can be implemented using kernel methods, such as

$$\hat{E}(h_i \mid w_i = (1-\tau)w_j, y_i = y_j, X_i = X_j) = \frac{\sum_{\substack{i=1 \\ i \in S_P}}^{n} h_i K\left(\frac{w_i - (1-\tau)w_j}{\lambda_n^w}\right) K\left(\frac{y_i - y_j}{\lambda_n^y}\right) I(X_i = X_j)}{\sum_{\substack{i=1 \\ i \in S_P}}^{n} K\left(\frac{w_i - (1-\tau)w_j}{\lambda_n^w}\right) K\left(\frac{y_i - y_j}{\lambda_n^y}\right) I(X_i = X_j)} \tag{2.9}$$

where $K(\cdot)$ denotes the kernel function, λ_n^w and λ_n^y are the smoothing (or bandwidth) parameters, and where, for convenience, X is assumed to take on a set of finite values. The kernel function and smoothing parameters satisfy standard assumptions that guarantee asymptotic consistency of the estimator.[13]

This example illustrates the feasibility of performing *ex ante* policy evaluation without having to introduce functional form or distributional assumptions. Although nonparametric, the method is not assumption-free. Indeed, the method requires an explicit characterization of a behavioral model and a number of key assumptions. These key assumptions ensure (1) that the hours of work function with the tax takes the same form as that without the tax, $\eta(w, y, \tau; X, \varepsilon) = \phi(\tilde{w}, y; X, \varepsilon)$, and (2) that variation in wage offers, like variation in the tax rate if such variation were available, is orthogonal to unobserved preference heterogeneity (conditional on y and X).

It is useful to consider how modifications in the theory affect nonparametric *ex ante* evaluation. One characteristic of the tax example is that the tax only entered the model through a change in the budget constraint. Suppose, instead, that the tax is also allowed to affect utility directly, that is $U = U(c, 1 - h, \tau; X, \varepsilon)$. In general, this modification would change the form of the hours of work function such that the tax rate would enter parametrically as a separate argument, that is, $\eta(w, y; \tau; X, \varepsilon) \neq \phi(\tilde{w}, y; X, \varepsilon)$. This utility specification might capture, for example, the utility value of public goods supplied through taxation or a direct "feel bad" effect from the existence of the tax. However, the condition that $\eta = \phi$ would be preserved if the utility function is additively separable in τ, $U(c, 1 - h; X, \varepsilon) + v(\tau)$, in which case the tax rate only affects hours of work through its effect on the net wage, $\tilde{w}$. Notice that the policy considered here, a wage tax, requires mandatory participation. As will be seen in a later example, additive separability is not sufficient if the policy allows for voluntary participation.

The orthogonality condition between preferences and wage offers is also restrictive. Individuals with stronger preferences for leisure over their life cycle would accumulate less work experience even in a myopic (static) model. In that case, wage offers, if affected by work experience, would be correlated with preference heterogeneity, if it was persistent, violating the conditional independence assumption. To mitigate this problem would require conditioning the estimator in equation 2.9 on work experience. More generally, X must include any variables that are correlated with both preferences for leisure and wage offers.

Partial Observability of Wages

The ability to perform nonparametric *ex ante* evaluation is also affected by the available data. Direct application of the matching estimator (or of a nonparametric estimator of the hours function, ϕ) is clearly not possible if wage offers are not observed for labor market nonparticipants. Absent their wage offers, a nonparticipant cannot be matched either to a participant or to another nonparticipant with the appropriate (after tax) wage offer.[14]

To perform an *ex ante* analysis requires that one be able to consistently estimate wage offers for the nonparticipants. A nonparametric approach would specify the wage function (in logs as is conventional and up to an additive error) as

$$\log w = \log w(z) + \xi, \tag{2.10}$$

where z are observable wage determinants, ξ are unobservable wage determinants, and $f(\xi \mid z) = f(\xi)$, but would make no assumptions about either the functional form of $w(z)$ or the distribution of ξ. A standard control function approach can be used without imposing parametric assumptions about the utility function if (as is discussed below) it is assumed that there is no preference heterogeneity (that is, dropping ε from equation 2.1).[15]

In that case, denoting the marginal rate of substitution function evaluated at zero hours of work as

$$M(y;X) = \frac{\partial U(y,1;X)/\partial l}{\partial U(y,1;X)/\partial c}$$

the participation rule is

$h = 0$ iff $M(y;X) \geq \log w(z) + \xi$,

$h > 0$ otherwise.

The probability that an individual with observables y, X, and z chooses to participate is

$$\Pr(h = 1 \mid y, X, z) = 1 - F_\xi(M(y,X) - \log w(z)),$$

which implies that

$$M(y,X) - \log w(z) = 1 - F_\xi^{-1}(\Pr(h = 1 \mid y, X, z)). \tag{2.11}$$

Now, the expected (log) wage for participants (the expected "accepted" log wage) is given by

$$E(\log w \mid h > 0) = \log w(z) + E(\xi \mid \xi > M(y,X) - \log w(z))$$

or

$$\begin{aligned} E(\log w \mid h > 0) &= \log w(z) + \frac{\int_{-\infty}^{\infty}\int_{M(y,X)-\log w(z)}^{\infty} \xi f(\xi)d\xi}{\int_{-\infty}^{\infty}\int_{M(y,X)-\log w(z)}^{\infty} f(\xi)d\xi} \\ &= \log w(z) + H(\Pr(h = 1 \mid y, X, z)). \end{aligned} \tag{2.12}$$

where the last equality follows from equation 2.11.[16] Thus,

$$\log w_{h>0} = \log w(z) + H(\Pr(h = 1 \mid y, X, z)) + u \tag{2.13}$$

where u has conditional mean zero by construction.

It is possible to identify the H function up to a constant if there is at least one continuous variable that affects the participation probability, $\Pr(h = 1 \mid y, X, z)$ that is not in z. Non-earned income, y, would satisfy that condition, as would any such variable in X (that is not in z). The identification argument proceeds as follows. Given a nonparametric estimate of $\Pr(h = 1 \mid y, X, z)$ from the data, one can identify H over its full support by fixing z and varying $\widehat{\Pr}(h = 1 \mid y, X, z)$ through variation in y and/or a continuous variable in X. Then, one can estimate $\log w(z)$ by varying z and, say, y such that H is held constant. However, the constant term in $\log w(z)$ cannot be separately identified from the constant term in H. To see this, simply redefine the H function as $H - h_0$ and add h_0 to equation 2.13. Similarly, redefine the $w(z)$ function as $w(z)/w_0$ and add $\log w_0$. Then the constant term in equation 2.13 is $\log w_0 + h_0$.

For some purposes, knowledge of the constant term is unnecessary, for example, in determining the effect of schooling or work experience on wage offers (the human capital stock). However, knowledge of the constant term is critical for the *ex ante* evaluation of the introduction of a wage tax. The matching estimator requires that one be able to compare the hours of work of an individual with wage offer w to an individual with wage offer $(1 - \tau)w$. Obviously, the wage offer of a nonparticipant cannot be identified without having an estimate of the constant term in $w(z)$.

It is thus not possible to be fully nonparametric if wage offers are not observed for nonparticipants. Suppose then that ξ is assumed to be normal but, as above, the form of $w(z)$ is left unspecified. In that case, equation 2.13 becomes

$$\begin{aligned} \log w_{h>0} &= \log w(z) + \sigma_\xi \frac{\varphi\left(\frac{H(\Pr(h = 1 \mid y, X, z))}{\sigma_\varepsilon}\right)}{1 - \Phi\left(\frac{H(\Pr(h = 1 \mid y, X, z))}{\sigma_\varepsilon}\right)} + u \\ &= \log w(z) + \sigma_\xi \lambda\left(\frac{H(\Pr(h = 1 \mid y, X, z))}{\sigma_\varepsilon}\right) + u \end{aligned} \tag{2.14}$$

where $\lambda(\cdot)$ is the familiar Mills ratio selection correction. Once the Mills ratio is constructed, $w(z)$ can be estimated without a confounding constant term.

Given an estimate of $w(z)$ and of σ_ξ (which is identified in equation 2.14 from variation in the Mills ratio), the matching estimator can be combined with simulation to obtain a consistent estimate of the policy effect. Suppose that the n observations in the data are ordered so that

the first n_1 are participants for whom wages are observed and the next $n-n_1$ observations are nonparticipants for whom wages are not observed. Consider a simulation in which $n - n_1$ values of the wage error, ξ, are drawn from $f(\xi) \sim N(0, \sigma_\xi)$, thus simulating a wage offer (using equation 2.10) for each nonparticipant given z.[17] Given that there is now a full set of wage offers, for nonparticipants from the simulation and for participants from the data, one can obtain for this simulation a matching estimator of the policy effect from equation 2.8 using equation 2.9. Performing $s = 1, \ldots, S$ such simulations and denoting the wage vector associated with the sth simulation as $w_s = (w_1, \ldots, w_{n_1}, w^s_{n_1+1}, \ldots, w^s_n)$ and the policy effect from equation 2.8 based on that wage vector as $\hat{\Delta}_s$, the estimated policy effect is

$$\hat{\hat{\Delta}} = \frac{1}{S}\sum_{s=1}^{S} \hat{\Delta}_s, \tag{2.15}$$

where the double hat indicates that, for finite S, equation 2.15 is an estimate of equation 2.8.

Measurement Error

It is often suspected that hours of work are measured with error. Under classical assumptions, with additive mean zero measurement error, the consistency property of the matching estimator is unaffected, although the precision of the estimator is reduced. Measurement error in hours often translates into measurement error in wages (for example, if wages are defined as earnings over some time period divided by hours over that same period). Following the argument in the preceding section, the existence of additive measurement error in (log) wages does not change the selection correction as given by equation 2.14. Measured wages do not affect the decision problem of the individual. Thus, as without measurement error, both $w(z)$ and σ_ξ are identified.[18] The matching estimator would be obtained by simulating wage offers for both those who work (based on their values of z) and, as above, for those who do not work. The policy effect, $\hat{\Delta}_s$, would be obtained for each set of simulation draws and used to obtain the estimator (equation 2.15).

Fixed Costs of Work

Among the earliest extensions of the labor supply model was the inclusion of fixed costs of work (see Cogan, 1981; Heckman and MaCurdy, 1982). Introducing a fixed money cost of work, ν_1, implies a budget constraint when there is no tax given by

$$C = wh + y - \nu_1 I(h > 0) \tag{2.16}$$

and the associated optimal hours function, $h^* = \varphi(w, y; \nu_1, X, \varepsilon)$. The budget constraint when there is a tax, in the presence of fixed costs, is now

$$\begin{aligned} C &= (1-\tau)wh + y - \nu_1 I(h > 0) \\ &= \tilde{w}h + y - \nu_1 I(h > 0). \end{aligned} \tag{2.17}$$

As seen from equations 2.16 and 2.17, the introduction of the fixed cost alters the budget constraint when there is a tax in the same way as when there is no fixed cost. Thus, as before, the same function, ϕ, governs the choice of hours with and without the tax. The matching estimator (equation 2.8 or 2.15) again recovers the policy effect.[19]

Child Care Costs

It is possible, also, within this framework to evaluate the effect of a child care subsidy on labor supply, say of a single female parent. If we let ς be the per-child hourly cost of child care, the budget constraint, ignoring taxes, equation 2.2, becomes

$$\begin{aligned} c &= wh - \varsigma n h + y \\ &= (w - \varsigma n)h + y \end{aligned}$$

where n is the number of children requiring child care when the mother works. Suppose the government provides a child care subsidy leading to a net per-child child care cost of ς'. Then, with the subsidy, the budget constraint is

$$c = (w - \varsigma' n)h + y.$$

By analogy to the wage tax, it is possible to evaluate the effect of the net (of subsidy) child care cost on hours by matching individuals with wage $w - \varsigma n$ to individuals with wage $\tilde{w} = w - \varsigma' n$. For any given market cost of child care, ς, the effect on hours of work of the subsidy $\varsigma - \varsigma'$ can be determined. Whether it is necessary to match also on n depends on how the utility function is specified, in particular on whether the marginal rate of substitution between leisure and consumption depends on n. One cannot be agnostic about the behavioral model. By reintroducing the wage tax, it is possible to estimate the joint effect of introducing a wage tax together with a child care subsidy on labor supply by a matching estimator with $\tilde{w} = (1-\tau)w - \varsigma' n$.

Nonlinear Taxes

Tax schedules are generally nonlinear. For that reason, there has developed a large methodological and empirical literature on the estimation of labor supply models in the presence of nonlinear taxes (see Blundell and MaCurdy, 1999, for an extensive review). Because the tax rate depends on the level of wage earnings, that is on wh, the existence of nonlinear taxes changes the budget constraint in a nontrivial way. In particular, equation 2.4 becomes

$$C = (1 - \tau(wh; \gamma))wh + y,$$

where the tax schedule, $\tau(wh; \gamma)$, depends on a vector of parameters, γ. Thus, the hours of work function is in general a function of the parameters that determine an individual's tax liability; with the utility function in equation 2.1, optimal hours are given by

$$h^{**} = \eta(w, y; \gamma; X, \varepsilon).$$

Contrary to the case of a proportional tax, the hours of work function in the presence of nonlinear taxes cannot be written as the same hours function that applies in the absence of taxes with a change in variables, that is, as $\phi(\tilde{w}, y; X, \varepsilon)$, where $\tilde{w}$ is a transform of w. The hours function is therefore not invariant to changes in the tax schedule and is, thus, not structural with respect to the tax. The implication is that starting from a world with no wage tax, it is not possible to perform an *ex ante* evaluation of the introduction of a nonlinear tax scheme.

However, recall that in the case of a proportional tax, the same *ex ante* evaluation could be performed when there is an existing proportional tax, τ_1, which has not varied in the past, that is, proposed to be changed to a new value τ_2. That result carries over, though, as will be seen in a more limited way, to the case of a change in an existing nonlinear wage tax. In several papers, Blomquist and Newey (1997, 2002) develop and implement a nonparametric estimation method for the usual case in which the tax schedule leads to a kinked budget constraint. They apply the method to an *ex post* evaluation of Swedish tax reforms making use of variation in tax schedules over time.[20] My interest is in showing the potential of this methodology for *ex ante* evaluation in the absence of variation in the tax schedule.

Figure 2.1 illustrates the budget constraint for a three-tier progressive tax schedule. In the figure, y_1 is non-earned income and $w_1 = w(1 - \tau_1)$ is the slope of the first segment, where w is the wage and τ_1 the tax rate over the first segment. At $h = k_1$, where earnings are $(1 - \tau_1)$

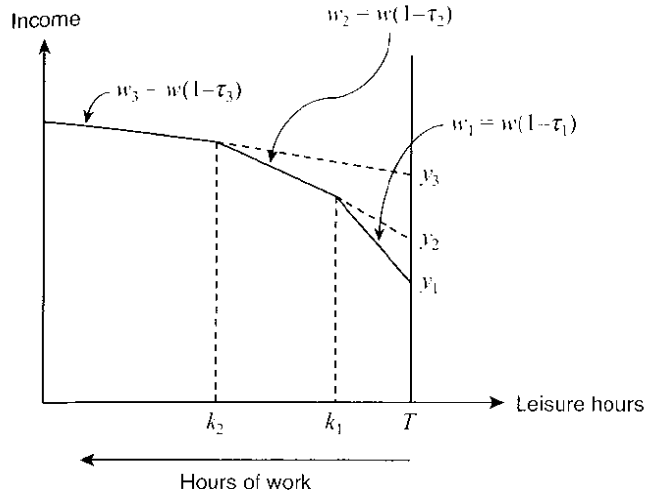

Figure 2.1
Progressive tax and labor supply.

wk_1, the marginal tax rate changes to τ_2. The second line segment thus has slope $w_2 = (1 - \tau_2)w$ with virtual income for that segment of y_2. At $h = k_2$, where earnings are $(1 - \tau_1)\, wk_1 + (1 - \tau_2)\, wk_2$, the marginal tax rate changes to τ_3. The third line segment thus has slope $w_3 = (1 - \tau_3)w$ with virtual income for that segment of y_3.

The insight of Blomquist and Newey (1997, 2002) is that under utility maximization, expected hours of work can be expressed as a function of the arguments of the budget constraint as just described. In the general case of J segments, optimal expected hours are given by the function

$$Eh^* = \phi(w_1, w_2, \ldots, w_J, y_1, y_2, \ldots, y_J, k_1, k_2, \ldots, k_{J-1}). \tag{2.18}$$

Estimating equation 2.18 nonparametrically, the goal of their analysis, is impractical for most actual tax systems because of the high dimensionality of the function. A major simplification in terms of feasibility is achieved because, as Blomquist and Newey show, equation 2.18 can be written as the sum of lower-dimensional objects. In the case where $J = 3$, as in the figure,

$$Eh^* = \phi_1(w_1, y_1, k_1) + \phi_2(w_2, y_2, k_1) + \phi_3(w_2, y_2, k_2) + \phi_4(w_3, y_3, k_2). \tag{2.19}$$

As they also note, because the kink points are functions of the wages and virtual incomes, namely

$$k_j = \frac{y_{j+1} - y_j}{w_j - w_{j+1}}, \tag{2.20}$$

the dimensionality of equation 2.19 can be reduced further. Accounting for equation 2.20,

$$Eh^* = \phi_{12}(w_1, y_1, w_2, y_2) + \phi_{34}(w_2, y_2, w_3, y_3) \tag{2.21}$$

which is of overall dimension 6 rather than 8 as in equation 2.19. The simplification in equation 2.21 is useful if the goal is to estimate the expected hours function. However, it is not useful for the more limited goal of *ex ante* policy evaluation.

In that context, I have taken as a premise that there has been no previous variation in the tax schedule. In that case, variation in the net wages w_1, w_2, and w_3 arises only from variation in the gross wage, w, and the expected hours function (equation 2.18) simplifies to

$$Eh^* = \phi(w, \mathbf{y}; \boldsymbol{\tau}) = \phi(\mathbf{w}, \mathbf{y}) \tag{2.22}$$

where $\mathbf{y}$ is the vector of virtual incomes, $\boldsymbol{\tau}$ is the vector of tax rates, and, in the second equality, $\mathbf{w}$ is the vector of *net* wages. Recall that in the case of a proportional tax, *ex ante* evaluation was possible over the "full" range of potential tax rates given enough wage variation in the data. In the case of a nonlinear tax, the set of policies that can be evaluated is limited. In particular, if the current tax regime is characterized by a tax schedule given by the vector $\mathbf{1} - \boldsymbol{\tau}$, then *ex ante* evaluation can be performed on alternative tax schedules given by a proportionate change in each element of $\mathbf{1} - \boldsymbol{\tau}$, that is, for tax schedules in which the ratios $\frac{1-\tau_i}{1-\tau_j}$ are unchanged for all i, j. The matching estimator of the effect of such a change in the tax schedule would be based on a comparison of the hours of work of individuals with gross wage w to individuals with gross wage βw, where β is either less than 1, representing a lower tax schedule, or greater than 1, representing a tax increase.[21]

The result that only prospective tax schedules that satisfy the constant ratio requirement can be evaluated clearly severely restricts the range of policies that can be evaluated. It is not possible to evaluate a policy that changes the tax rate along only one (or multiple) segment while holding the other tax rates constant.[22] That implies that it is also not possible to evaluate new schedules that have either a smaller or larger number of segments than the existing tax schedule. Variation in the tax schedule, as used in the Blomquist and Newey (2002) application, allows for the evaluation of more types of schedules, although the set of policies is still restricted to those that satisfy the constant-ratio property of each schedule. If the variation in tax schedules is small, extrapolations outside of the constant ratio set of taxes can only be achieved through a parametric assumption.[23]

Parametric estimation of labor supply models with kinked budget constraints is considerably more complicated when the budget constraint has nonconvexities.[24] With nonconvexities there may be multiple tangencies between indifference curves and budget segments, which implies that to determine optimal hours of work requires making global comparisons of utility levels. That issue does not arise in the *ex ante* evaluation problem; implementing the matching estimator assumes only that hours observations are optimally chosen and does not require that the optimization problem be solved.

Applications

There have not been any applications, to my knowledge, of the matching estimator in the *ex ante* evaluation of tax policy, although there have been a few *ex post* evaluations using the Blomquist and Newey (2002) nonparametric approach. These include studies of Swedish tax reform in the 1980s by Blomquist, Eklof, and Newey (2001) and Blomquist and Newey (2002), the study of the U.S. Tax Reform Act of 1986 by Kumar (2008), and the study of policy changes in the U.S. EITC, AFDC, and Food Stamp programs during the 1990s by Wu (2005). All of these papers exploit policy variation. None use their estimates to perform counterfactual (*ex ante*) policy evaluation exercises. I take up applications of the matching estimator in the example that follows.

2.1.1.2 *School Attendance Subsidy*

Many developing countries have adopted conditional cash transfer (CCT) programs designed to increase human capital investment and reduce poverty. Such programs have been tried or currently exist in Argentina, Brazil, Chile, Colombia, Egypt, Guatemala, Honduras, Indonesia, Jamaica, Malawi, Mexico, Nicaragua, Panama, Turkey, and Zambia. CCT programs of this kind have usually included a subsidy for regular school attendance.

A policy maker contemplating the introduction of a CCT program would presumably prefer to have an estimate of the effectiveness of the policy, including how the effect of the subsidy would vary with the size and structure of the subsidy, prior to implementation. Having such

an estimate would enable the policy maker to calculate the costs and benefits of alternative subsidy schedules and thus to make an informed decision about the best structure of the subsidy to implement. As in the case of the wage tax, it is possible to perform an *ex ante* evaluation without variation in the price (tuition) of schooling and without parametric assumptions. Also as with the wage tax, the feasibility of nonparametric *ex ante* evaluation is dependent on the theory that is posited about how households make school attendance choices. I again begin with the simplest optimization model and then make modifications to explore extensions and limitations of the approach.

The Basic Framework

Consider a household with one school-age child deciding between having the child attend school or having the child work in the labor market. School attendance provides direct utility to the household; work provides income, and thus consumption, to the household. Letting y denote household income net of the child's earnings, w the child's wage offer, $s \in \{0, 1\}$ school attendance, and X observable and ε unobservable taste shifters, the parents choose s to maximize

$$U(C, s; X, \varepsilon) \tag{2.23}$$

subject to the budget constraint

$$C = y + w(1 - s). \tag{2.24}$$

The optimal school attendance choice is

$$s^* = \phi(y, w; X, \varepsilon), \tag{2.25}$$

where in equation 2.25, $s^* = \phi(y, w; X, \varepsilon) = 1$ if $U(y, 1; X, \varepsilon) > U(y + w, 0; X, \varepsilon)$; that is, where the utility to the household from the child attending school exceeds that from the child working, and $s^* = \phi(y, w; X, \varepsilon) = 0$ otherwise.

Suppose that the government introduces a conditional cash transfer program that provides a subsidy of τ if the child attends school. Under this policy, the household faces a new budget constraint given by

$$C = y + w(1 - s) + \tau s. \tag{2.26}$$

Noting that one can rewrite equation 2.26 as

$$C = (y + \tau) + (w - \tau)(1 - s),$$

the optimal choice of s can be written as the same function as without the subsidy, namely

$$s^{**} = \phi(y + \tau, w - \tau; X, \varepsilon), \tag{2.27}$$

where $\phi(y + \tau, w - \tau; X, \varepsilon) = 1$ if $U(y + \tau, 1; X, \varepsilon) > U(y + w, 0; X, \varepsilon) = U((y + \tau) + (w - \tau), 0; X, \varepsilon)$ and $\phi(y + \tau, w - \tau; X, \varepsilon) = 0$ otherwise.[25] As seen in equation 2.27, the subsidy can be viewed as reducing the child wage by the amount of the subsidy, that is, reducing the opportunity cost of school attendance, and at the same time increasing household income by the amount of the subsidy. The effect of what amounts to an equal reduction in w and increase in y increases the utility from having the child attend school but leaves the utility from having the child work (not attend school) unchanged.

The implication of equation 2.27 is that a household with income y, child wage w, and preference shifters X and ε that receives the subsidy will make the same schooling decision as a household with income $\tilde{y} = y + \tau$, child wage $\tilde{w} = w - \tau$, and the same preference shifters that did not receive the subsidy. Assuming that ε is independent of w and y conditional on X, that is, $f(\varepsilon \mid w, y, X) = f(\varepsilon \mid \tilde{w}, \tilde{y}, X) = f(\varepsilon \mid X)$, and that child wage offers are observed for all households, the effect of the subsidy on school attendance can be obtained by a similar matching procedure as for the wage tax example, namely

$$\hat{\Delta} = \frac{1}{n} \sum_{\substack{j=1 \\ j,i \in S_P}}^{n} \hat{E}(s_i \mid w_i = w_j - \tau, y_i = y_j + \tau, X_i = X_j) - s_j(w_j, y_j, X_j) \tag{2.28}$$

where $s_j(w_j, y_j, X_j)$ is the school attendance decision of household j with characteristics (w_j, y_j, X_j) (that is, without the subsidy) and $\hat{E}(s_i \mid w_i = w_j - \tau, y_i = y_j + \tau, X_i = X_j)$ is an estimate of the expected school attendance that household j would choose if faced with the subsidy. As before, the average can only be taken over the region of overlapping support, S_p, that is, over the set of households j whose values $w_j - \tau$ and $y_j + \tau$ lie within the observed support of wages w_i, household income y_i, and observable characteristics X_i. The first term in equation 2.28 can be estimated from a nonparametric regression of s_i on w_i and y_i evaluated at the points $w_i = w_j - \tau$, $y_i = y_j + \tau$, $X_i = X_j$ (see the discussion below).

It is useful to consider modifications in the model to understand the robustness of the method. Throughout it is assumed that wage offers are observed for both school nonattendees (children who work) and school attendees (children who do not work). Partial observability of wages raises no new issues in this context than have already been discussed in the wage tax example.

However, that is not the case if the household's utility is directly affected by participation in the subsidy program. Unlike the wage tax, participation in the school attendance subsidy program is voluntary; indeed, the take-up rate is isomorphic to the policy impact. In that case, the school attendance decision will necessarily depend on the extent to which the policy induces a "feel good" or "feel bad" effect. Welfare programs are, for example, thought to have a stigma effect, which reduces participation. Without a strong prior, presumably based on external evidence about the existence and magnitude of such effects, it is impossible to perform a valid *ex ante* policy evaluation. It would, however, be possible to evaluate the impact of a change in an existing policy if the utility or disutility of the policy were independent of the policy change, for example, if the (dis)utility associated with a subsidy of τ were the same as that with a subsidy of a multiple or fraction of τ.[26]

Multiple Children—Exogenous Fertility

The method can be extended to households with more than one school- (and work-) age child. Consider a subsidy program in which, as has been the case in actual programs, the subsidy amount varies with grade level. Thus, the subsidy faced by a household with multiple children is a schedule that depends on the grade levels that their children are eligible to attend. Denote $n_k = 1$ if there is a child in the household who would be attending grade level $k = 1, \ldots, K$ and $n_k = 0$ otherwise.[27] The household is assumed to obtain utility from the attendance of each child according to

$$U(C, s_1 n_1, s_2 n_2, \ldots, s_K n_K; X, \boldsymbol{\varepsilon}), \tag{2.29}$$

where $\boldsymbol{\varepsilon} = (\varepsilon_1 n_1, \ldots, \varepsilon_K n_K)$ is a vector of child-specific school attendance preference "shocks." The household budget constraint in the presence of the subsidy schedule is

$$\begin{aligned} C &= y + \sum_{k=1}^{K} w_k (1 - s_k) n_k + \sum_{k=1}^{K} \tau_k s_k n_k \\ &= \left(y + \sum_{k=1}^{K} \tau_k n_k \right) + \sum_{k=1}^{K} (w_k - \tau_k)(1 - s_k) n_k \end{aligned} \tag{2.30}$$

where τ_k is the subsidy level for a child attending grade level k. Maximizing expression 2.29 subject to equation 2.30 yields the school attendance demand function for a child of grade level k

$$s_k^{**} = \phi_k \left(y + \sum_{k=1}^{K} \tau_k n_k, (w_1 - \tau_1) n_1, (w_2 - \tau_2) n_2, \ldots, (w_K - \tau_K) n_K; X; \boldsymbol{\varepsilon} \right)$$
$$\text{for } k = 1, \ldots, K. \tag{2.31}$$

As seen in equation 2.31, the school attendance demand function for any child depends not only on the wage and subsidy level for that child but also on the wage and subsidy levels of all other children in the household. Thus, the matching procedure in the case of multiple children requires that the matched households have the same number of children and the same configuration of children in terms of their grade levels.[28] In that case, families with income y, wages $w_1 n_1, \ldots, w_k n_k$, and observable characteristics X would be matched with families with the same value of X, income $y + \sum_{k=1}^{K} \tau_k n_k$, and wages $(w_1 - \tau_1) n_1, (w_2 - \tau_2) n_2, \ldots, (w_K - \tau_K) n_K$.

Multiple Children—Endogenous Fertility

In the previous case, when fertility was taken to be exogenous, accommodating multiple children required a straightforward change in the matching procedure. When fertility is a choice, even if wages and subsidy levels are the same for all children, matching on the number of children is no longer a valid procedure. The reason is that the number of children is now itself a function of w, y, and the observable and unobservable preference shifters. Varying w or y while holding the number of children constant implies that the unobservables must also change. Thus, $f(\varepsilon | w, y, n, X) \neq f(\varepsilon | n, X)$, which violates a necessary condition for the validity of the matching procedure.

More fundamentally, when fertility is a choice, the demand functions for school attendance are no longer invariant to the subsidy. To accommodate fertility choice, assume that the utility function is augmented to include the number of children and a preference unobservable that shifts the marginal utility of the number of children.[29] In addition, to isolate the issue raised by the endogeneity of fertility, assume that neither the wage nor the school attendance subsidy varies with child characteristics (grade level). Then, denoting n as the number of children and distinguishing between the school attendance unobservable (vector) as $\boldsymbol{\varepsilon}_s$ and the fertility unobservable as ε_n, the utility function is

$$U(C, s_1, \ldots, s_n, n; X, \boldsymbol{\varepsilon}_s, \varepsilon_n)$$

where s_i is the school attendance of child i, and the budget constraint given the subsidy (after rearranging) is

$$C = (y + n\tau) + (w - \tau)\sum_{i=1}^{n}(1 - s_i).$$

Utility maximization leads to the demand system

$$s_i^{**} = \eta_i(y, w - \tau; \tau; X, \varepsilon_s, \varepsilon_n), \text{ for } i = 1, \ldots, n^{**}, \tag{2.32}$$

where

$$n^{**} = \eta^n(y, w - \tau; \tau; X, \varepsilon_s, \varepsilon_n).$$

Note that the subsidy, τ, enters as a separate parameter in the demand functions; one cannot write equation 2.32 as the same function that arises without the subsidy evaluated at different values of w and y.[30] Thus, the nonparametric approach is not applicable with endogenous fertility.[31]

Life Cycle—Perfect Foresight

It is useful to consider the extent to which nonparametric *ex ante* evaluation can be extended to a life cycle context. I take up first the case in which the household is assumed to have perfect foresight and, for expositional ease, consider a household that will have only one child (and knows it). The household makes a school attendance decision for that child in each of $t = 1, \ldots, T$ periods, corresponding to the ages of the child for which work is legal. The household is assumed to have perfect foresight about household income, child wages, preference shifters, and the time sequence of school attendance subsidy levels each period. Household utility is

$$U(C_1, \ldots, C_T, s_1, s_2, \ldots, s_T; \mathbf{X}, \boldsymbol{\varepsilon}), \tag{2.33}$$

where $\mathbf{X} = (X_1, \ldots, X_T)$ and $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_T)$. The household, under the subsidy policy, is assumed to satisfy the period-by-period budget constraint (no savings or borrowing)

$$\begin{aligned} C_t &= y_t + w_t(1 - s_t) + \tau_t s_t \\ &= (y_t + \tau_t) + (w_t - \tau_t)(1 - s_t). \end{aligned}$$

The optimal school attendance decision for each age, t, is determined by a global comparison of utilities at all 2^T possible combinations of s_t over all $t = 1, \ldots, T$.

Defining $\widetilde{y(\tau)} = (y_1 + \tau_1, y_2 + \tau_2, \ldots, y_T + \tau_T)$ and $\widetilde{w(\tau)} = (w_1 - \tau_1, w_2 - \tau_2, \ldots, w_T - \tau_T)$, the school attendance demand function in each period is thus

$$s_t^{**} = \phi_t(\tilde{y}, \tilde{w}; \mathbf{X}, \boldsymbol{\varepsilon}) \text{ for } t = 1, \ldots T.$$

The estimation of the subsidy effect requires matching untreated households (of given $\mathbf{X}$) with income profile $\widetilde{y(0)}$ and child wage offer profile $\widetilde{w(0)}$ to households with income profile $\widetilde{y(\tau)}$ and child wage offer profile $\widetilde{w(\tau)}$.[32]

Dynamics—Imperfect Foresight

To simplify the exposition, assume that there are only two schooling periods and, as above, that there can be no borrowing or saving. The budget constraint under the subsidy policy is

$$C_t = (y_t + \tau) + (w_t - \tau)(1 - s_t) \text{ for } t = 1, 2,$$

where the subsidy is assumed to be constant over the two periods and, in the case of period 2, independent of attendance in period 1. The school attendance decision in any period t depends on a comparison of the remaining expected lifetime utility of the household under the two alternatives, $s_t = \{0, 1\}$. Given the utility function (expression 2.33) with $T = 2$ (and dropping $\mathbf{X}$ for convenience), the schooling decision in period 2 is

$$\begin{aligned} s_2^{**} &= 1 \text{ iff } U(y_2 + \tau, s_1, 1; \varepsilon_2) - U((y_2 + \tau) + (w_2 - \tau), s_1, 0; \varepsilon_2) \geq 0 \\ &= 0 \text{ otherwise.} \end{aligned}$$

Note that the schooling decision in period 2 depends on whether or not the child attended school in period 1 because utility is not intertemporally separable in school attendance. If ε_2 is unknown in period 1 and the unobserved preference shifter is serially independent, $f(\varepsilon_2 \mid \varepsilon_1) = f(\varepsilon_2)$, s_1 will not be correlated with ε_2.[33] In that case, conditional on s_1, the matching procedure as in the static model can be used to estimate the impact of the subsidy on period 2 school attendance. The overall impact is then the weighted sum of the conditional (on s_1) impacts, with the weights the proportion of households choosing each s_1.

The schooling decision in period 1 is based on a comparison of the sum of the expected discounted utilities over the two periods under each of the two school attendance alternatives. With these denoted as $V(s)$, they are given by

$$V(s_1 = 1) = U(y_1 + \tau, 1; \varepsilon_1) + \tag{2.34}$$

$$\delta E \max(U(y_2 + \tau, 1, 1; \varepsilon_2), U((y_2 + \tau) + (w_2 - \tau), 1, 0; \varepsilon_2)),$$

$$V(s_1 = 0) = U(y_1 + \tau) + (w_1 - \tau), 0; \varepsilon_1) + \tag{2.35}$$

$$\delta E \max(U(y_2 + \tau, 0, 1; \varepsilon_2), U((y_2 + \tau) + (w_2 - \tau), 0, 0; \varepsilon_2)),$$

where δ is the discount factor and the expectation is taken over whatever elements inside the E max function are unknown to the household and thus can be viewed as random from the perspective of the household. The school attendance decision in period 1 is thus

$$s_1^{**} = 1 \text{ iff } V(s_1 = 1) - V(s_1 = 0) \geq 0$$
$$= 0 \text{ otherwise.}$$

The feasibility of *ex ante* evaluation depends on what elements are in the household's information set. For example, and perhaps the least realistic, suppose that the household has perfect foresight about the period 2 income and child wage but not about the unobserved preference shifter, ε_2. In that case the school attendance demand function is given by

$$s_1^{**} = \phi_1(y_1 + \tau, y_2 + \tau, w_1 - \tau, w_2 - \tau; \varepsilon_1, f(\varepsilon_2)).$$

As in the perfect foresight case, the same school attendance function governs the choice of s_1 with and without the subsidy, and the same matching procedure can be adopted in estimating the policy effect. Of course, the attendance function differs from the perfect foresight case in that it depends on the density of ε_2 as opposed to the value of ε_2. Note that the evaluation of the subsidy policy does not entail estimating the ϕ_1 function (or the underlying utility function and preference distribution) and so does not discriminate a model in which all households have perfect foresight from one in which they have imperfect foresight.

Consider instead the case in which the household is uncertain about its future income, y_2, but otherwise has perfect foresight. Further assume that household income is an *iid* process. Integrating over the density of y_2 in equations 2.34 and 2.35 changes the form of the school attendance demand function. In particular, τ will in general enter parametrically, so that

$$s_1^{**} = \eta_1(y_1 + \tau, w_1 - \tau, w_2 - \tau, \tau; \varepsilon_1, \varepsilon_2, f(y_2)). \tag{2.36}$$

The matching estimator based on the household income and child wage comparisons fails in this case.[34] Given the form of equation 2.36 and the presumed population invariance of the subsidy schedule, nonparametric *ex ante* evaluation is not possible. Clearly, the same result obtains if the household is uncertain about the period 2 child wage.

Child Home Production

Children in developing countries may engage in home production as an alternative to attending school or working in the labor market. To

capture that fact, consider again the static single-child model in which the household values the child "leisure" option, that is, where the child neither attends school nor works in the market for a wage. Assuming that school attendance, work, and leisure are mutually exclusive alternatives and letting $l = \{0, 1\}$ indicate the leisure option, the household utility function[35] is now

$$U(C, s, l; \varepsilon_s, \varepsilon_l),$$

and the budget constraint under the subsidy is

$$C = y + w(1 - s - l) + \tau s. \tag{2.37}$$

With equation 2.37 rewritten as

$$C = (y + \tau) + (w - \tau)(1 - s - l) - \tau l, \tag{2.38}$$

the school attendance demand function can be seen to depend not only on $y + \tau$ and $w - \tau$, as in the model without the leisure option, but also parametrically on τ, that is,

$$s^{**} = \eta(y + \tau, w - \tau, \tau; \varepsilon_s, \varepsilon_l). \tag{2.39}$$

The school attendance demand function without the subsidy, $s^{**} = \phi(y, w; \varepsilon_s, \varepsilon_l)$, thus differs from that with the subsidy, which implies that the matching estimator cannot be used to perform an *ex ante* evaluation of the subsidy policy.[36]

Interestingly, and perhaps also surprisingly, the matching estimator based on the model with child leisure corresponds to a different policy, one that instead of subsidizing school attendance provides a subsidy to households in which the child does not work in the market, that is, in which a subsidy is provided if the child either attends school or stays home. Under that policy, the budget constraint becomes

$$C = y + w(1 - s - l) + \tau(s + l)$$
$$= (y + \tau) + (w - \tau)(1 - s - l.)$$

Thus, the school attendance and leisure demand functions,

$$s^{**} = \phi^s(y + \tau, w - \tau; \varepsilon_s, \varepsilon_l)$$

and

$$l^{**} = \phi^l(y + \tau, w - \tau; \varepsilon_s, \varepsilon_l),$$

are the same functions that would apply without the subsidy policy ($\tau = 0$).

The matching estimator of the subsidy effect on school attendance that compares households with income y and child wage w to households with income $y + \tau$ and child wage $w - \tau$ is thus consistent with either of two models and two corresponding policies. Inference about the effect of the school subsidy policy relies on there being no leisure option that the household values. There is otherwise a fundamental ambiguity in the policy interpretation of the matching estimator.

2.1.2 Structural Parametric Approach

2.1.2.1 *School Attendance Subsidy*

There are many models of behavior, as shown in the preceding discussion, for which a nonparametric approach to *ex ante* policy evaluation is not conceptually possible. And, although I have not explicitly discussed the computational feasibility of that approach, it is clear that empirical tractability can become an issue for models that are more complex, that is, where the matching variables are of high dimension. I now consider how parametric assumptions address these problems.

The Basic Framework

Like the nonparametric approach, the parametric approach also requires that the researcher adopt a particular behavioral model. However, the parametric approach makes additional assumptions about functional forms and error distributions. The value of these extratheoretic auxiliary assumptions is that they provide restrictions that enable the identification of the structural parameters of the model that are necessary for *ex ante* policy evaluation in cases where the nonparametric approach fails. To fix ideas, consider the single-child static school attendance decision model given by equations 2.23 and 2.24 in which the utility function (equation 2.23) takes the specific parametric form[37]

$$U(C,s;\varepsilon) = C + \alpha s + \beta C s + \varepsilon s, \tag{2.40}$$

and where ε, which shifts the marginal utility of school attendance, is assumed to be normally distributed in the population, $\varepsilon \sim N(0,\sigma_\varepsilon^2)$.[38] The school attendance decision under the subsidy policy is given by

$$\begin{aligned} s &= 1 \text{ iff } \varepsilon \geq (w-\tau) - \alpha - \beta(y+\tau), \\ &= 0 \text{ otherwise.} \end{aligned}$$

And the probability of attendance is given by

$$\Pr(s=1) = 1 - \Phi\left(\frac{(w-\tau) - \alpha - \beta(y+\tau)}{\sigma_\varepsilon}\right),$$

where Φ is the standard cumulative normal.[39] Note that, consistent with the discussion of the model in the nonparametric case, the attendance probability is a function of $(y + \tau)$ and $(w - \tau)$. The parameters of the model can be estimated by maximum likelihood from preprogram data alone, $\tau = 0$, provided there is variation in child wage offers, assumed to be observed both for children who work and for children who attend school, and variation in household income. A normalization of σ_ε, as is usual in the case of a probit, is unnecessary because of the implicit normalization of utility in monetary units.[40] Given parameter estimates, the effect of introducing a subsidy of τ on the attendance rate can be calculated from

$$\hat{\Delta}_s = \Phi\left(\frac{(w-\tau) - \hat{\alpha} - \hat{\beta}(y+\tau)}{\hat{\sigma}_\varepsilon}\right) - \Phi\left(\frac{w - \hat{\alpha} - \hat{\beta} y}{\hat{\sigma}_\varepsilon}\right)$$

for any hypothetical subsidy level.[41] Unlike the nonparametric case, there is no condition on the support of $y + \tau$ and $w - \tau$.

As previously shown, the ability to do nonparametric *ex ante* evaluation of the subsidy policy is lost if child leisure is added to the household utility function. To see how parameterizing the model allows *ex ante* evaluation in this case, consider adding child leisure to equation 2.40 as follows:

$$U(C,s,l;\varepsilon) = C + \alpha s + \beta C s + \gamma l + \varepsilon_s s + \varepsilon_l l, \tag{2.41}$$

where $l \in \{0, 1\}$, s and l are mutually exclusive, and ε_l is an unobservable that alters the marginal utility of child leisure. The two preference shifters are assumed to be jointly normally distributed with mean zero and variance-covariance matrix $\Lambda = \begin{pmatrix} \sigma_{\varepsilon_s}^2 & \cdot \\ \sigma_{\varepsilon_s\varepsilon_l} & \sigma_{\varepsilon_l}^2 \end{pmatrix}$. The budget constraint is given, as in equation 2.38.

It is useful to write the utilities associated with the three mutually exclusive choices, school attendance (s), leisure (l), and work (h), where $s + l + h = 1$:

$$U(s=1) = (1+\beta)(y+\tau) + \alpha + \varepsilon_s,$$

$$U(l=1) = y + \gamma + \varepsilon_l,$$

$$U(h=1) = y + w.$$

This choice model is thus in the form of a trivariate probit with the utility of one of the alternatives a nonstochastic normalized function of observables. The model parameters are identified (see Heckman and Sedlacek, 1985) and can be estimated from preprogram data ($\tau = 0$) with variation in w and y.[42] Given parameter estimates, the effect of introducing a subsidy of τ on school attendance (and on child employment) can be calculated from the cumulative bivariate normal distribution of (ε_s, $\varepsilon_s - \varepsilon_t$) for any hypothetical subsidy level. Note that $s = 1$ requires that

$$\beta(y + \tau) + \tau + \alpha - \gamma + \varepsilon_s - \varepsilon_t \geq 0$$

and

$$\beta(y + \tau) - (w - \tau) + \alpha + \varepsilon_s \geq 0,$$

which implies that the school attendance function depends, as shown in the discussion of the nonparametric matching estimator, not only on $y + \tau$ and $w - \tau$, but also separately on τ.

The structural parametric approach has been used to evaluate the PROGRESA program in two papers (Todd and Wolpin, 2006; Attanasio, Meghir, and Santiago, in press). Both of these applications adopt DCDP models, although the actual models differ greatly in their behavioral assumptions. There have been a number of recent surveys of the DCDP approach (Aguiregebaria and Mira, 2010; Todd and Wolpin, 2010; Keane, Todd, and Wolpin, 2011). Nevertheless, a brief description of the development of DCDP models will help to place these applications in context. The introduction of DCDP models is associated with independent contributions by Gotz and McCall (1984), Miller (1984), Pakes (1986), Rust (1987), and Wolpin (1984). The basic insight that enabled the estimation of DCDP models is that any DCDP model can be cast as a static discrete choice estimation problem.

To illustrate that insight, first consider the simple static model in which the only decision is school attendance of a single child. Assume that wages are observed (by the researcher) only for children who work. To focus on essentials, the utility function is specified as additively separable in consumption and school attendance (thus eliminating the income effect). Household i at time t maximizes

$$U(C_{it}, s_{it}; \varepsilon_{it}) = C_{it} + \alpha_{it} s_{it},$$

where preferences are time varying according to $\alpha_{it} = X_{it}\beta + \varepsilon_{it}$, subject to a budget constraint

$$C_{it} = y_{it} + w_{it}(1 - s_{it})$$

and a wage offer function

$$w_{it} = Z_{it}\gamma + \eta_{it}. \tag{2.42}$$

In anticipation of the dynamic setting, it is useful to distinguish the household's state space, Ω_{it}, consisting of all of the determinants of the household's decision, that is X_{it}, ε_{it}, Z_{it}, η_{it}, from the part of the state space available to the researcher, Ω_{it}^-, consisting only of the determinants X_{it}, Z_{it}. Denoting v_{it}^* (the latent) as the difference in the utility of the household under the alternatives $s_{it} = 1$ and $s_{it} = 0$,

$$\begin{aligned} v_{it}^* &= Z_{it}\gamma - X_{it}\beta + \eta_{it} - \varepsilon_{it} \\ &= \xi^*(\Omega_{it}^-) + \xi_{it}, \end{aligned} \tag{2.43}$$

where ε and η are distributed joint normal with covariance matrix Λ.[43] The sample likelihood incorporating the wage information for working children is

$$L(\theta; X_{it}, Z_{it}) = \Pi_{i=1}^{I-1} \Pr(s_{it} = 1, w_{it} \mid \Omega_{it})^{s_{it}} \Pr(s_{it} = 0 \mid \Omega_{it}^-)^{1-s_{it}}$$

where

$$\Pr(s_{it} = 1, w_{it} \mid \Omega_{it}^-) = \Pr(\xi_{it} \geq -\xi_{it}^*(\Omega_{it}^-),$$
$$\eta_{it} = w_{it} - Z_{it}\gamma),$$
$$\Pr(s_{it} = 0 \mid \Omega_{it}^-) \mid = \Pr(\xi_{it} < -\xi_{it}^*(\Omega_{it}^-)).^{44}$$

The parameters to be estimated include β, γ, σ_ξ^2, σ_η^2, and $\sigma_{\varepsilon\eta}$. As is well known, joint normality is sufficient to identify the wage parameters (γ and σ_η^2) as well as $(\sigma_\eta^2 - \sigma_{\varepsilon\eta}) / \sigma_\xi$ without exclusion restrictions (Heckman, 1979). The data on work choices identify γ/σ_ξ and β/σ_ξ. To identify σ_ξ, note that there are three possible types of variables that appear in the likelihood function, variables that appear only in Z, that is, only in the wage function, variables that appear only in X, that is, only in the utility function, and variables that appear in both X and Z. Having identified the parameters of the wage function (the γs), the identification of σ_ξ (and thus also $\sigma_{\varepsilon\eta}$) requires the existence of at least one variable that appears only in the wage equation. For example, variables that affect the demand for child labor, and thus the child wage, must not affect the utility value the household places on the child's school attendance.

In the static model there is no connection between the decision made in the current period and future utility. Thus, even if agents are forward looking, maximizing the expected present value of discounted lifetime utility would be equivalent to maximizing current utility in each period. There are many ways in which dynamics may arise in the model. For

example, suppose the child's wage increases with actual work experience, H. In that case, rewrite equation 2.42 as

$$w_{it} = Z_{it}\gamma_1 + \gamma_2 H_{it} + \eta_{it}, \tag{2.44}$$

where $H_{it} = \sum_{\tau=1}^{\tau=t-1}(1-s_{i\tau})$ is work experience at the start of period t. Given this specification, working (not attending school) in any period increases future wages, which, if the household is forward looking, will reduce the incentive for having the child attend school. Alternatively, or in addition, we could suppose that the child wage and/or parents' utility depends not only on current attendance but also on the current number of years of schooling the child has completed. In that case, attending school in any period affects future utility either indirectly through its effect on future wage offers or directly on future utility.

Continuing with the example, modified to account for the wage offer function in equation 2.44, assume that the couple maximizes the expected present discounted value of remaining lifetime utility at each period starting from an initial period, $t = 1$, and ending at period T, the assumed terminal decision period. If we let $V_t(\Omega_{it})$ be the maximum expected present discounted value of remaining lifetime utility at t given the state space and discount factor δ (the value function),

$$V_t(\Omega_{it}) = \max_{s_{it}} E\left\{\sum_{\tau=t}^{\tau=T}\delta^{\tau-t}[U_{i\tau}^1 s_{i\tau} + U_{i\tau}^0(1-s_{i\tau})] \,|\, \Omega_{it}\right\},$$

where U^1 and U^0 are the levels of utility with $s = 1$ and $s = 0$. The state space at t consists of all factors known to the household at t that affect current utility or the probability distribution of future utilities. With the wage relationship given by equation 2.44, H_t becomes part of the state space and evolves according to

$$H_{it} = H_{i,t-1} + (1 - s_{i,t-1}).$$

The value function can be written as the maximum over the two alternative-specific value functions, $V_t^k(\Omega_{it})$, $k \in \{0, 1\}$,

$$V_t(\Omega_{it}) = \max(V_t^0(\Omega_{it}), V_t^1(\Omega_{it})), \tag{2.45}$$

each of which obeys the Bellman equation

$$\begin{aligned} V_t^k(\Omega_{it}) &= U_{it}^k + \delta E[V_{t+1}(\Omega_{i,t+1}) \,|\, \Omega_{it}, s_{it} = k] \text{ for } t < T, \\ &= U_{iT}^k \text{ for } t = T. \end{aligned} \tag{2.46}$$

The expectation in equation 2.46 is taken over the distribution of the random components of the state space at $t + 1$ conditional on the state space elements at t, that is, over the unconditional joint distribution of the random shocks at $t + 1$, given that all shocks are mutually serially independent.

The latent variable in the dynamic case is the difference in alternative specific value functions, $V_t^1(\Omega_{it}) - V_t^0(\Omega_{it})$, namely

$$\begin{aligned} v_t^*(\Omega_{it}) &= Z_{it}\gamma_1 + \gamma_2 H_{it} - X_{it}\beta - \varepsilon_{it} + \eta_{it} \\ &\quad + \delta\{E[V_{t+1}(\Omega_{i,t+1}) \,|\, \Omega_{it}, s_{it} = 1] - E[V_{t+1}(\Omega_{i,t+1}) \,|\, \Omega_{it}, s_{it} = 0]\} \\ &= \xi_{it}^*(\bar{\Omega}_{it}) + \xi_{it}. \end{aligned} \tag{2.47}$$

When the latent variable function in the dynamic case (equation 2.47) is compared to that of the static case (equation 2.43), the only difference is the appearance of the difference in the future component of the expected value functions under the two alternatives in equation 2.47. A full solution of the dynamic programming problem consists of finding $E[\max(V_t^0(\Omega_{it}), V_t^1(\Omega_{it}))]$ at all values of $\bar{\Omega}_{it}$, denoted by $E\max(\bar{\Omega}_{it})$, for all $t = 1, \ldots T$.

Estimation of the dynamic model requires that the researcher have data on the children's work experience, H_{it}. More generally, assume that the researcher has longitudinal data and denote t_{1i} and t_{Li} as the first and last periods of data observed for household i. In that case, the likelihood function is

$$L(\theta; X_{it}) = \prod_{i=1}^{I}\prod_{\tau=t_{1i}}^{\tau=t_{Li}} \Pr(s_{i\tau} = 0, w_{i\tau} \,|\, \bar{\Omega}_{it})^{1-s_{i\tau}} \Pr(s_{i\tau} = 1 \,|\, \bar{\Omega}_{it})^{s_{i\tau}},$$

where

$$\begin{aligned} &\Pr(s_{i\tau} = 0, w_{i\tau} \,|\, \bar{\Omega}_{it}) = \Pr(\xi_{i\tau} \geq -\xi_{i\tau}^*(\bar{\Omega}_{i\tau}), \\ &\eta_{i\tau} = w_{i\tau} - Z_{i\tau}\gamma_1 - \gamma_2 H_{it}), \\ &\Pr(s_{i\tau} = 1 \,|\, \bar{\Omega}_{it}) = 1 - \Pr(\xi_{i\tau} \geq -\xi_{i\tau}^*(\bar{\Omega}_{i\tau})).^{45} \end{aligned}$$

Identification requires the same exclusion restriction as in the static case, that is, the appearance of at least one variable in the wage equation, that is, in Z, that does not affect the utility value of school attendance, that is, does not appear in X. Work experience, H_{it}, would serve that role if it did not also enter into X_{it}. Note that the difference in the future component of the expected value functions under the two alternatives in equation 2.47 is a nonlinear function of the state variables, Z_{it}, H_{it}, and X_{it} and depends on the same set of parameters as in the static case. Equation 2.47 can be rewritten as

$$v_t^*(\Omega_{it}) = Z_{it}\gamma_1 + \gamma_2 H_{it} - X_{it}\beta + \delta G(Z_{it}, H_{it}, X_{it}) - \varepsilon_{it} + \eta_{it}, \tag{2.48}$$

where $G(\cdot)$ is the difference in the future component of the expected value functions; the nonlinearities in G that arise from the distributional

and functional form assumptions may be sufficient to identify the discount factor. Estimation of the dynamic model is in principle no different from the estimation of the static model. The practical difference is that the dynamic programming problem must be solved at each iteration of the likelihood function, and the solution of the dynamic programming problem is often not analytic.

Analyzing the impact of the policy experiment of introducing the school attendance subsidy is not any different in the dynamic than in the static setting. If the critical value of ξ that determines the choice of working or attending school is $\xi_{it}^{*}(\Omega_{it}^{-})$ without the subsidy, with the subsidy it is $\xi_{it}^{*}(\Omega_{it}^{-})-\tau$. The increase in school attendance is thus $\Phi(\frac{-\xi_{it}^{*}(\Omega_{it}^{-})+\tau}{\sigma_{\xi}})-\Phi(\frac{-\xi_{it}^{*}(\Omega_{it}^{-})}{\sigma_{\xi}})$, again highlighting the importance of having the necessary exclusion restrictions to identify σ_{ξ}.

The Multinomial Choice Problem

The dynamic analogue to the static multinomial choice problem is conceptually no different than it was for the binary choice problem. Indeed, it does not do much injustice to simply allow the number of mutually exclusive alternatives, and thus, the number of alternative-specific value functions in equation 2.45 to be greater than two. Analogously, if there are $K > 2$ mutually exclusive alternatives, there will be $K-1$ latent variable functions (relative to one of the alternatives, arbitrarily chosen). The static multinomial choice problem raises computational issues with respect to the calculation of the likelihood function because its calculation requires multivariate integrations. Having to solve the dynamic multinomial choice problem, that is, for the E max one can use $(V^{0}(\Omega_{it}), V_{t}^{1}(\Omega_{it}), \ldots, V_{t}^{K}(\Omega_{it}))$ function that enters the multinomial version of equation 2.45 at all values of Ω_{it}^{-} and at all t, adds significantly to that computational burden. A number of methods have been developed to ameliorate this burden (see Keane, Todd, and Wolpin, 2011, for a review). For example, Rust (1987) shows that in the case of additive errors that are distributed independent extreme value for each mutually exclusive alternative, the E $\max_t$ functions have closed forms.

Unobserved Heterogeneity

The stochastic components of the model, preference, and wage shocks were assumed to be mutually serially uncorrelated. There is nothing that rules out general serial correlation other than computational feasibility in solving the dynamic programming problem and in estimating the model. A standard specification that allows for serial dependence, conditional on observable state variables, assumes that agents can be distinguished, in terms of preferences and opportunities, by a fixed number of types. For example, if a household was of type j, the preference for school attendance might be specified as $\alpha_{ijt} = \alpha_{oj} + X_{it}\beta + \varepsilon_{it}$, and the child's wage offer as $w_{ijt} = \gamma_{oj} + Z_{it}\gamma_{1} + \gamma_{2}H_{it} + \eta_{it}$.[46] A type j household would be distinguished by the $(\alpha_{oj}, \gamma_{oj})$ pair. Thus, potentially, families who value child schooling more might also have children who are more (or less) productive in the labor market. The dynamic program, in this case, must be solved for each of the J types, and the likelihood function is a weighted average of the type-specific likelihoods. The weights are the proportions of each type in the sample and are estimated along with the other parameters.[47]

More Flexible Specifications

Estimable DCDP models are not nonparametrically identified (Rust, 1994) and usually fully parametrically specified.[48] However, any DCDP that can be numerically solved can, in principle, be estimated, and the DCDP structure does not restrict the choice of functional forms for preferences, technologies, and institutional constraints (e.g., tax rules), including the way in which unobservables enter (e.g., nonadditive errors, serial correlation), nor does it restrict the distributional assumptions of unobservables. The restrictions that are typically imposed arise from practical considerations, for example, about the size of the state space and the number of parameters that must be estimated, as well as from parameter identification.

Alternative Estimation Approaches

The main limiting factor in estimating DCDP models is the computational burden associated with the iterative process, regardless of whether estimation is by maximum likelihood or method of moments. It is therefore not surprising that there have been continuing efforts to reduce the computational burden of estimating DCDP models. Hotz and Miller (1993) developed a method for implementing DCDP models that does not involve solving the DP model, that is, calculating the $E\max(\Omega_{it}^{-})$ functions. They prove that, for additive errors, the $E\max(\Omega_{it}^{-})$ functions can be written solely as functions of conditional choice probabilities and state variables for any joint distribution of additive shocks. More recently, Imai, Jain, and Ching (2009) and Norets (2009) have developed computationally practical Bayesian approaches to the solution and estimation of DCDP models that rely on Markov Chain Monte Carlo (MCMC) methods.

2.2 Applications

2.2.1 Nonparametric *Ex Ante* Evaluation with the Matching Estimator

In this section, I report on two applications of the matching estimator to school attendance subsidy programs. Todd and Wolpin (2008) conducted an *ex ante* evaluation of the Mexican PROGRESA program, and Azevedo, Bouillon, and Yanez-Pagans (2009) evaluated the Mexican *Oportunidades* program (which replaced and extended PROGRESA).

The PROGRESA program, introduced in Mexico in 1998, provided transfers to households contingent on their children's regular school attendance.[49] To evaluate the impact of the program, the Mexican government conducted a randomized experiment: of 506 rural villages involved in the experiment, 320 were randomly assigned to receive the treatment and the remaining 186 to serve as controls. Eligibility of households residing in villages chosen to receive the treatment was based on a "marginality" index that depended on preprogram characteristics such as whether the home had a dirt floor, ownership of assets, household composition, and children's school attendance.

The subsidy schedule, as shown in table 2.1 for the first semester of the 1998/1999 academic year, depends on the child's grade level and gender.[50] The subsidy begins at grade 3 and extends through grade 9.

Table 2.1
Monthly School Attendance Subsidies

		PROGRESA[a]		*Oportunidades*[b]	
School Level	Grade	Girls	Boys	Girls	Boys
Primary	3	70	70	120	120
	4	80	80	140	140
	5	105	105	180	180
	6	135	135	240	240
Secondary	7	210	200	370	350
	8	235	210	410	370
	9	235	225	450	390
Upper Secondary	10	0	0	675	585
	11	0	0	715	630
	12	0	0	760	665

[a]Todd and Wolpin (2008).
[b]Azevedo, Bouillon, and Yanez-Pagans (2009).

The subsidy level increases with grade level, recognizing the higher opportunity cost of school attendance for older children (children are legally permitted to work in the formal labor market at age 12), and is slightly higher for girls at secondary school grades (grades 7–9), who traditionally have lower school enrollment rates at those grades. Overall, the transfers received by the treated households were substantial, amounting to about 20–25 percent of total annual income (Skoufias and Parker, 2000).

In addition to data on school attendance, which is necessary for the experimental evaluation, detailed information was also collected from all treatment and control village households on household demographics, school attainment of household members, household income, and employment and wages of children. Data were gathered in two baseline surveys, conducted in October 1997 and March 1998, and several follow-up surveys, in October 1998, May 1999, and November 1999. Supplemental data were also gathered at the village level, including distance to the nearest secondary school, distance to the nearest city, and a village-level minimum wage for day laborers. The experiment ran for 2 years, after which households in the control villages were incorporated into the program. Data were collected on both eligible and ineligible households in the village. There were approximately 9,000 households in the control villages and 15,000 in the treatment villages.

The PROGRESA program was renamed the *Oportunidades* program in 2002 and extended school attendance subsidies to urban areas and to upper secondary grade levels (grades 10–12). Unlike the PROGRESA program, the extension was not implemented experimentally. There were other differences as well. In the rural program, a census of the targeted villages was conducted, and all households that met the eligibility criteria for the program were informed of their eligibility status. For cost reasons this type of census was not feasible in urban areas, and an alternative system of advertised sign-up periods was adopted. Households that thought themselves potentially eligible had to apply at local offices that were placed in areas with high concentrations of poor households. Table 2.1 shows the subsidy schedule for the first semester of the 2006/2007 academic year. As seen, the schedule is similar in structure to the rural program, with the subsidy increasing with grade level and differing by gender in secondary and upper-secondary grades. The subsidy amounts differ in nominal terms, although they are similar after adjustment for changes in the national

price level. A baseline survey had been conducted in the fall of 2002, just prior to the start of the program. The sample respondents consisted of eligible and ineligible households in the treatment areas and of households in nontreatment areas that would meet the eligibility criteria. Two additional rounds of data were collected in the fall of 2003 and 2004.

2.2.1.1 *PROGRESA*

Todd and Wolpin (2008) used a sample of children aged 12 to 15 in 1998 residing in *control* group villages who were reported to be children of the household head and for whom information was available in the 1997 and 1998 surveys. Household income was taken to be the sum of earned income of the children's parents, and the child wage was the reported village level minimum wage, which was taken as measuring the wage (offer) available to all children in the village. Both eligible and ineligible households were used in order to facilitate matching on household income. This information was available for roughly half the villages in the sample. The minimum monthly child wage ranged from 330 to 1,320 pesos, with a median of 550 pesos, and monthly household income from 8 to 13,750 pesos, with a median of 660 pesos. Todd and Wolpin (2008) estimated the impact of the PROGRESA subsidy program on school enrollment using two modeling frameworks previously discussed, one that treats each child in the household as a singleton and one that accounts for multiple children (exogenous fertility) in the household. The estimation method allows the school enrollment decision to potentially differ for girls and boys, which would accommodate, for example, differences in the utility that parents get from girls' and boys' schooling.

For the single-child model, the estimator of the predicted program effect is given by

$$\hat{\Delta}_s = \frac{1}{n}\sum_{\substack{j=1 \\ j,i\in S_P}}^{n} \{\hat{E}(s_i \mid w_i = w_j - \tau_j, y_i = y_j + \tau_j, g_i = g_j) - s_j(w_j, y_j, g_j)\},$$

where g_j denotes the child's gender and where τ_j is the subsidy level for which the child is eligible, which, as noted, varies by grade level. This estimator matches program-eligible control-group children with offered wage w_j and household income y_j to other (program-eligible and -ineligible) control group children with offered wage $w_j - \tau_j$ and

$y_i = y_j + \tau_j$, with the matches restricted to be between children of the same gender.

Expected school attendance, $\hat{E}(s_i \mid w_i = w_j - \tau_j, y_i = y_j + \tau_j, g_i = g_j)$, is estimated nonparametrically using a kernel regression estimator. With $w_0 = w_j - \tau_j$ and $y_0 = y_j + \tau$, the estimator is given by

$$\hat{E}(s_i \mid w_i = w_0, y_i = y_0, g_i = g_0) = \frac{\sum_{\substack{i=1 \\ i\in S_P}}^{n} s_i K\left(\frac{w_i - w_0}{\lambda_n^w}\right) K\left(\frac{y_i - y_0}{\lambda_n^y}\right) 1(g_i = g_0)}{\sum_{\substack{i=1 \\ i\in S_P}}^{n} K\left(\frac{w_i - w_0}{\lambda_n^w}\right) K\left(\frac{y_i - y_0}{\lambda_n^y}\right) 1(g_i = g_0)},$$

where $K(\cdot)$ denotes the kernel function and λ_n^w and λ_n^y are the smoothing (or bandwidth) parameters.[51]

The nonparametric estimator is defined only at points where the data density is positive. For this reason, estimation is restricted to points of evaluation that lie within the region S_P, where $S_P = \{(w, y) \in R^2\}$ such that $f(w, y) > 0$) and $f(w, y)$ is the density. It was determined empirically whether a particular point of evaluation (w_0, y_0) lies in S_P by estimating the density at each point and checking whether it lies above a cutoff trimming level, q_α, that is small and positive, that is, whether $\hat{f}(w_0, y_0) > q_\alpha$, where $\hat{f}(\cdot,\cdot)$ is a nonparametric estimate of the density. The cutoff level q_α corresponds to the 2 percent quantile of the positive estimated density values.[52]

For the multiple-child case, Todd and Wolpin (2008) consider the potential earnings of all children in the household. If all the children within a household had the same subsidy levels and potential wages, the estimator for the program effect would be given by:

$$\hat{\Delta}_s = \frac{1}{n}\sum_{\substack{j=1 \\ j,i\in S_P}}^{n} \{\hat{E}(s_i \mid w_i = w_j - \tau_j, y_i = y_j + n\tau_j, n_i = n_j, g_i = g_j) - s_j(w_j, y_j, n_j, g_j)\}, \quad (2.49)$$

where n_j denotes the number of children in child j's household and where matches are restricted to households with the same numbers of children. This estimator needs to be slightly modified to take into account that different children within the same household face different potential subsidies. To accommodate this feature, Todd and Wolpin (2008) estimate the program effect from

$$\hat{\Delta}_s = \frac{1}{n}\sum_{\substack{i=1 \\ j,i\in S_P}}^{n} \{\hat{E}(s_i \mid w_i = w_j - \bar{\tau}_j,\ y_i = y_j + n\bar{\tau}_j,\ n_i = n_j,\ g_i = g_j) - s_j(w_j, y_j, n_j, g_j)\}.$$

where $\bar{\tau}_j$ is the average subsidy level of the children in the household. The child's wage offer, which is the village-level wage, does not vary for children within households.[53]

Table 2.2 reports program impacts using the matching estimator based on the single- and multiple-child models. Estimates are reported separately over three different age ranges by gender and for both girls and boys. The estimation results that combine boys and girls of different age ranges still restrict matches to be between children of the same gender and the same age bracket. That is, a girl aged 12–13 would only be matched to other girls in the same age range, even for the results that aggregate across categories.[54] The percentage of observations that lie within S_P is also shown.

As seen in table 2.2, all of the treatment effect estimates for both the single- and multiple-child models are positive, although nothing in the implementation of the matching estimator ensures that the estimated

Table 2.2
Ex Ante Impact of PROGRESA Subsidies on School Attendance by Age and Gender[a]

	Single-Child Model		Multiple-Child Model	
	Impact	% Overlapping Support	Impact	% Overlapping Support
Girls				
Age 12–13	0.01 (.03)[b]	87%	0.05 (.03)[b]	68%
Age 14–15	0.01 (.04)	83%	0.09 (.05)	61%
Age 12–15	0.06 (.03)	86%	0.06 (.03)	64%
Boys				
Age 12–13	0.06 (.03)	91%	0.04 (.04)	67%
Age 14–15	0.07 (.05)	89%	0.11 (.06)	68%
Age 12–15	0.06 (.03)	90%	0.07 (.04)	68%
Girls and Boys				
Age 12–13	0.04 (.02)	89%	0.04 (.03)	67%
Age 14–15	0.09 (.04)	86%	0.10 (.04)	64%
Age 12–15	0.06 (.02)	88%	0.07 (.02)	66%

[a]Todd and Wolpin (2008).
[b]Bootstrap standard errors in parentheses.

ex ante treatment effect will have that sign. Within-gender estimates, however, for a number of age ranges have 95 percent confidence intervals that span zero. Estimates that combine girls and boys have confidence intervals that, except in one case, do not span zero. The matching estimator implies that the PROGRESA program increased school attendance of girls aged 12–15 by 6 percentage points regardless of the model and that of boys aged 12–15 by 6 or 7 percentage points, depending on the model.

In considering the impact of any conditional cash transfer program, it is desirable to know to what extent the conditionality makes a difference and whether similar impacts might be achieved through unconditional transfers. Therefore, Todd and Wolpin (2008) used a matching estimator to determine whether giving families an unconditional income transfer in the amount of 5,000 pesos per year would significantly impact school enrollments, an amount almost half of parental income. Their estimates, based on matching households by parent income levels (households with income y and households with income $y + 5000$), indicate that the unconditional income transfer would not lead to any statistically significant impacts on school enrollment.

2.2.1.2 *Oportunidades*

Azevedo, Bouillon, and Yanez-Pagan (2009) implemented the matching estimator of the single-child model (equation 2.49) using data from a large biannual nationally representative household survey, the Mexican National Household Income and Expenditure Survey, rather than the much smaller data set collected as part of the *Oportunidades* program. The 2006 national data set contains over 9 million children aged 12–18 and identifies beneficiaries of the *Oportunidades* program. Moreover, unlike *Oportunidades* program data, which are mainly composed of poor households, the national data include a large sample of higher-income households as is useful for the matching procedure. Matching is conditioned on household size and on the number of children over age 14 but not on gender or age subbrackets. The sample consists of households that are not beneficiaries of the program; households in this sample that would be eligible for inclusion as beneficiaries are matched potentially to those that would not be eligible. The matching estimator implies that the *Oportunidades* program increased the school attendance rate of eligible 12- to 18-year-olds by 1.7 percentage points.[55]

Both Todd and Wolpin (2008) and Azevedo, Bouillon, and Yanez-Pagan (2009) performed counterfactual exercises of alternative subsidy

schedules. These additional *ex ante* evaluations are discussed below. I first turn to a discussion of applications of the structural parametric approach.

2.2.2 Structural Parametric *Ex Ante* Evaluation

In this section, I report on two structural parametric evaluations of the PROGRESA school attendance subsidy program based on DCDP models, those by Todd and Wolpin (2006) and by Attanasio, Meghir, and Santiago (in press). Todd and Wolpin (2006) performed an *ex ante* evaluation using control group households, whereas Attanasio, Meghir, and Santiago (in press) conducted an *ex post* evaluation using both control and treatment group households. Both, however, performed *ex ante* evaluations of alterations in the program.

It is important to stress that although both papers adopt the DCDP approach, the models differed nontrivially in their structure. We present each of the models in some detail both to illustrate those differences and, for those not familiar with the estimation of DCDP models, to highlight implementation methods and their associated assumptions.

2.2.2.1 *The Todd and Wolpin Model*

The structure of the Todd and Wolpin (2006) model is an extended version of the models previously presented. In describing the model, I therefore forego writing the complete model equations to avoid additional notation. In the Todd and Wolpin model, in each year of their (known) finite lifetimes, a married couple decides on whether each of their children between the ages of 6 and 15 will attend school, remain at home, or, for those aged 12 to15, work in the labor market. These alternatives are assumed to be mutually exclusive. The couple also decide whether the wife will become pregnant in that period (while the woman is fecund).[56]

The couple receives flow utility in each period from consumption, from their current stock of children, from their children's current years of schooling (for example, the average years of schooling of their children, the number of children with 9 years of schooling), from the set of children at home by their ages and gender, and from the wife's pregnancy status. There are also preference interactions, for example, consumption is interacted with the number of children, the number of children at home, and their average schooling. To capture a quality-quantity tradeoff (Becker and Lewis, 1973; Rosenzweig and Wolpin, 1980), the number of children is interacted with average years of schooling completed. Additional interactions allow for complex choice dynamics. For example, the value the parents place on an older girl remaining at home may depend on the existence of pre-school-age children, and reentering school may entail a psychic cost. There is also a utility cost to attending school (grades 7–9) that depends on the distance from the village to a school. Households differ in their preferences over choice variables according to their discrete "type," and household preferences are subject to time-varying serially independent shocks.

Household income includes the income of both parents and the wage income of the children who work, both of which depend on the discrete household type and are subject to time-varying serially independent shocks. A child's wage (offer) also depends on the child's age and sex and the distance of the village to the nearest city. Attendance in school is not the same as grade completion. Grade progression is probabilistic, given attendance, and depends on the grade the child is attending, the child's age and gender, and the household type. In each period, household consumption is equal to household income.

It is useful to demonstrate the mechanisms through which the subsidy may have an effect, particularly because the estimation is conducted without recourse to data on the treatment villages. Consider the school attendance decision in the static model with household utility given by equation 2.40. As seen in that model, the introduction of a subsidy of τ increases the probability of school attendance by the same amount as would a reduction in the child wage together with an increase in household income by the same subsidy amount, τ. Absent an income effect, as in the utility specification equation 2.40 with $\beta = 0$, the subsidy effect could be identified solely from the wage effect, exactly as in the wage tax example. That is not the case in the Todd and Wolpin (2006) model, as there are several avenues through which there are income effects.[57]

First, utility is constant relative risk aversion CRRA in consumption, which in itself leads to an income effect. Second, although in the Todd and Wolpin (2006) specification consumption is separable from contemporaneous school attendance, it is not separable from past school attendance (years of schooling). A forward-looking model, as in Todd and Wolpin (2006), that implies that households that draw income from a distribution with a higher mean, for example, will also choose to send children to school more often; that is, there is a "permanent" income effect on school attendance. Given that the subsidy is multiperiod, a

larger subsidy in every period will also increase school attendance through an income effect. To see this, consider a two-period perfect-foresight model, as previously discussed, in which the flow utilities in the two periods are:

$$U_2 = C_2 + \alpha s_2 + \gamma C_2 s_1 + \varepsilon_2 s_2,$$

$$U_1 = C_1 + \alpha s_1 + \varepsilon_1 s_1$$

The budget constraint in each period is $C_t = (y + \tau) + (w - \tau)s_t$, where both income and the child wage are assumed to be the same in both periods. Note that in this specification there will be no contemporaneous income effect on the choice of school attendance in either period. Thus, there is no income effect on period 2 school attendance. The period 1 choice is determined by comparing the discounted sum of utilities over the two periods under the two alternatives, which are given by

$$V_1(s_1 = 1) = (y + \tau) + \alpha + \varepsilon_1 + \delta \max\{(y + \tau)(1 + \gamma) + \alpha + \varepsilon_2, (y + w)(1 + \gamma)\},$$

$$V_1(s_1 = 0) = (y + w) + \delta \max\{(y + \tau) + \alpha + \varepsilon_2, y + w\}.$$

The difference, $V_1(s_1 = 1) - V_1(s_1 = 0)$, can take on any of four different forms depending on γ and ε_2. For example, if $\gamma < 0$, it is possible that when $s_1 = 1$ it is optimal to choose $s_2 = 1$, and when $s_1 = 0$, it is optimal to choose $s_2 = 0$. In that case, the decision rule in period 1 is

$$s_1 = 1 \text{ iff } -(w - \tau)(1 + \delta) + \delta\gamma(y + \tau) + \alpha(1 + \delta) + \varepsilon_1 + \delta\varepsilon_2 \geq 0$$
$$= 0 \text{ otherwise.}$$

Thus, with forward-looking behavior, $\delta > 0$, the effect of the subsidy on period 1 school attendance depends on household income (discounted one period). The conclusion can be shown to hold when there is imperfect foresight about future preferences, where ε_2 is not known in period 1 (see below).

In addition to income effects, the subsidy effect will not be equivalent to the wage effect alone whenever the subsidy itself also affects the school attendance decision. Recall that in the static model in which child leisure is added to household utility, the subsidy level affects the school attendance decision (see equation 2.39) differently than the child wage even if there is no income effect ($\beta = 0$). As this discussion shows, the utility specification can be quite general without requiring the use of treatment data to be able to perform the *ex ante* evaluation of introducing the attendance subsidy.[58] As already noted, and discussed in more detail below, control data alone are not sufficient if there is a direct utility effect of participating in the program.

The parameters of the Todd and Wolpin (2006) model are estimated by simulated maximum likelihood. The observed outcomes at each period are: (1) the choice (from the feasible set) made by the couple of whether or not to have a pregnancy, which children to send to school, which to work in the market, and which to remain at home; (2) the wages received by the children who work in the market; (3) the success or failure of those children who attend school to complete a grade level; and (4) parental income. The likelihood jointly incorporates the entire set of outcomes. Implementation requires additional assumptions. In particular, initial conditions at the time of marriage, the ages of marriage of both parents and the distances to the city and school, are assumed to be exogenous conditional on a household's type. For households observed in the first survey year (in 1997) that already have school-age children, the assumption of serial independence in the preference, child wage, and parent income shocks implies that the state variables at any time (for example, the stock of children, the average schooling of children) are also exogenous conditional on household type. Given that household type is unobserved, the likelihood function is a weighted average of the type-specific likelihoods. The weights, the type probabilities, are specified as derived from a multinomial logit in the state variables.[59] This procedure can be justified as an approximation to the type probability functions that would be the outcome of solving the dynamic programming problem back to the "true" initial period (say, the date of marriage) and using Bayes' rule to update the type probabilities given outcomes up to the first observation period of each household.[60] In addition, because the child wage shock is household specific, having an observation on the wage for two children in the same household working in the same period who have different wages (conditional on the relevant observable determinants of child earnings, child age, and sex) would lead to a degenerate likelihood. To avoid this degeneracy, Todd and Wolpin (2006) assume that the children's wages are measured with error.[61] As previously discussed, identification of the model parameters, in addition to functional form and distributional assumptions, requires an exclusion restriction, that there be at least one variable that affects the child wage that does not affect preferences. Todd and Wolpin (2006) assume that distance to the nearest city satisfies that restriction.

The model parameters enter the likelihood through the choice probabilities that are computed from the solution of the dynamic programming problem. Subsets of parameters enter through other structural relationships as well, such as child wage offer functions, the parents' income function, and the school failure probability function. The estimation procedure, maximizing the likelihood function, involves iterating between the solution of the dynamic program and the calculation of the likelihood.

The estimation sample consisted of landless nuclear households in which there was a woman under the age of 50 reported to be the spouse of the household head. Estimation was based on 1,316 households that were in the control villages. As of the October 1997 survey, there were 4,012 children born to these control households and 1,958 children between the ages of 6 and 15.[62] In estimating the model, to avoid a choice-based sampling problem, Todd and Wolpin (2006) used data on both program-eligible and -ineligible households as the eligibility criteria depended on the number of children attending school, a choice variable in the model.

As is standard practice in the DCDP literature, Todd and Wolpin (2006) compute goodness-of-fit statistics for a number of dimensions of the data. I report one such table because it illustrates the rationale for the seemingly complex specification of the utility function.[63] Table 2.3 compares the actual and predicted school attendance rates for children whose schooling attainment differs from their maximum potential, defined as the level they could have achieved had they enrolled at age 6 and attended school continuously without repeating grades. Note that school attendance rates fall precipitously as children aged 12 to 15

Table 2.3
Actual and Predicted School Attendance Rates by Number of Years Lagging Behind in School, Age 12–15[a]

	Boys			Girls		
Age	Actual	Predicted	χ^2	Actual	Predicted	χ^2
Not behind	88.3	82.1	8.50	83.8	78.2	6.02
Behind 1 year	79.8	76.4	1.56	75.4	74.5	0.09
Behind 2 years	65.8	62.5	0.91	52.9	51.0	0.20
Behind 3 years or more	49.1	51.7	0.62	44.7	42.7	0.39

$\chi^2(.05, 1) = 3.84$
[a]Todd and Wolpin (2006).

fall further behind in school, with the rate falling from 88.3 percent for boys who are not behind to 49.1 percent for boys 3 or more years behind. The decline predicted by the model is somewhat less steep, from 82.1 percent to 51.7 percent. The model fit is rejected for children not behind but not for children 1, 2, or 3 or more years behind. Without incorporating a utility cost of school reentry that varied with how much children deviated from their maximum potential schooling, the model would not have been able to mimic the patterns shown in table 2.3.[64]

This table illustrates another point. It is probably obvious that the utility function was not specified in its final form prior to beginning the estimation. In fact, the final specification was the outcome of a process of trial and error in which specifications were modified (many times) as the fit of the model was assessed. Thus, the final fit statistics reflect this data-mining process and cannot be taken as a pure indication of the validity of the model. My best guess is that all DCDP models mature in this way. Todd and Wolpin (2006) deal with this by using the treatment sample for out-of-sample validation. I defer discussion of this methodology, that is, of holding out a portion of the sample (in this case the entire treatment sample) for model validation, until the concluding section of this chapter.

2.2.2.2 *The Attanasio, Meghir, and Santiago Model*

The Attanasio, Meghir, and Santiago (2011) model differs in fundamental ways from the model in Todd and Wolpin (2006). Attanasio, Meghir, and Santiago (2011) consider the binary choice of school or work, excluding the "at home" option, and the decision about each child is considered independently of the other children in the household. Because the model differs from those considered in the previous discussion and the authors raise the issue of whether or not to hold out part of the sample, as in Todd and Wolpin (2006), I provide the details of the model specification. I maintain the variable notation previously used as much as possible.

Attanasio, Meghir, and Santiago (2011) specify the flow utility of a household associated with the school attendance of child i at age t as:

$$U_{it}^s = Y_{it}^s + \theta_s \tau_{it} D_{it} \tag{2.50}$$

$$Y_{it}^s = \mu_i^s + X_{it}\gamma^s + \psi^s S_{it} + Q_{1it}\varsigma_1 I(S_{it} \leq 5) + Q_{2ij}\varsigma_2 I(6 \leq S_{it} \leq 8) + \varepsilon_{it}^s, \tag{2.51}$$

where Y_{it}^s, the net benefit of attending school (absent the subsidy) for child i at age t, is composed of a child-specific fixed factor μ_i^s, preference

shifters (X_{it}), the current school attainment of the child (S_{it}), factors that affect the cost of attending primary school (Q_{1it}), factors that affect the cost of attending secondary school (Q_{2it}), child- and time-varying factors observable to the household but unobservable to the researcher, ε_{it}^s, and $I(\bullet)$ is the indicator function equal to one when the expression inside the parenthesis is true and equal to zero otherwise. Recall that τ_{it} is the subsidy amount, where the i subscript is included to capture the difference in the subsidy for children of different genders and where child age is implicitly treated as identical to grade level (which is what actually determines the subsidy amount). The school attendance subsidy, τ_{it}, affects the utility of school attendance for eligible households only in the treatment villages, with $D_{it} = 1$ denoting a treatment village and $D_{it} = 0$ denoting a control village

The flow utility from work is given by

$$U_{it}^w = Y_{it}^w + \theta_w w_{it} \tag{2.52}$$

$$Y_{it}^w = \mu_i^w + X_{it}\gamma^w + \psi^w S_{it} + \varepsilon_{it}^w, \tag{2.53}$$

where, in addition to the wage, w_{it}, the utility from working includes a child-specific fixed factor μ_i^w, preference factors (X_{it}), the current school attainment (S_{it}), and child- and time-varying factors observable to the household but unobservable to the researcher, ε_{it}^w. The unobservables, ε_{it}^s and ε_{it}^w, are assumed to be independently distributed extreme values. The child-specific fixed factor is modeled as discrete types.

It is instructive to consider the static version of the model. If $Y_{it}^s = \overline{Y_{it}^s} + \varepsilon_{it}^s$ and $Y_{it}^w = \overline{Y_{it}^w} + \varepsilon_{it}^w$, then the school attendance decision would be determined by

$$\begin{aligned} s_{it} &= 1 \text{ iff } \varepsilon_{it}^s - \varepsilon_{it}^w \geq -(\overline{Y_{it}^s} - \overline{Y_{it}^w}) - \theta_s \tau_{it} D_{it} + \theta_w w_{it} \\ &= 0 \text{ otherwise.} \end{aligned} \tag{2.54}$$

Attanasio, Meghir, and Santiago (2011) stress the point that the model allows the effect of the subsidy θ_s to differ from the effect of the child wage, θ_w. They argue that θ_s cannot be identified without variation in $\tau_{it}D_{it}$, that is, without the treatment data, thus making it impossible to perform an *ex ante* evaluation of the introduction of the subsidy policy based solely on control group data.

Although the necessity of combining treatment and control data seems transparent from equation 2.54, the validity of that conclusion depends on the underlying constrained optimization problem of the household from which the attendance decision (equation 2.54) is

derived. Consider the previous static model in which equation 2.40 is augmented to include a direct effect of the subsidy on household utility, that is, where utility is given by

$$U(C, s; \varepsilon) = \theta_w C + \alpha s + \beta C s + \lambda \tau s + \varepsilon s.^{65} \tag{2.55}$$

The direct utility effect, as seen in equation 2.55, is λ per unit of subsidy. With the budget constraint under the subsidy given as $C = y + w(1 - s) + \tau s$, the school attendance decision is

$$\begin{aligned} s &= 1 \text{ iff } \varepsilon \geq \theta_w w - \alpha - \beta y - (\theta_w + \beta + \lambda)\tau, \\ &= 0 \text{ otherwise.} \end{aligned} \tag{2.56}$$

It is easily seen that equation 2.56 is identical to equation 2.54 with $\beta y = \overline{Y_{it}^s} - \overline{Y_{it}^w}$ and $\theta_s = \theta_w + \beta + \lambda$, so the two models are observationally equivalent. The first thing to notice is that even if $\lambda = 0$, so that there is no direct utility effect of participating in the program, θ_s will not equal θ_w as long as $\beta \neq 0$, that is, as long as there is an income effect. In that case, if $\lambda = 0$, $\theta_s = \theta_w + \beta$ can in fact be identified using only control group data, for which $\tau_{it}D_{it} = 0$ given that θ_w and β are each identified.[66] Indeed, as previously shown, interpreting the Attanasio, Meghir, and Santiago (2011) model as above, *ex ante* evaluation can be conducted nonparametrically.

The second thing to notice is that the Attanasio, Meghir, and Santiago (2011) specification of the attendance decision (equation 2.54), even when using treatment group data, does not allow separate identification of β and λ. To obtain identification, the constrained optimization problem that corresponds to the specification of the alternative specific utility functions (equations 2.50–2.53) would need to be specified.[67] In that way, parameter restrictions, as in equation 2.56, can be exploited. Otherwise, the models with $\lambda = 0$ and $\lambda \neq 0$ are observationally equivalent, and estimation provides no evidence on the existence of a direct utility effect. Thus, although they are not explicit, by asserting that treatment data are necessary, Attanasio, Meghir, and Santiago (2011) are assuming that $\lambda \neq 0$. In their model, the need for treatment data would arise solely from the existence of a direct utility effect of participation.[68]

An important feature of the Attanasio, Meghir, and Santiago (2011) model, not incorporated by Todd and Wolpin (2006), is the inclusion of equilibrium effects of the subsidy program on the village-level child wage. Because eligible households are a significant proportion of the village population, the withdrawal of children from the labor market

due to the subsidy program will tend to increase the equilibrium child wage in the village. This equilibrium effect on the wage will mitigate the impact of the subsidy program on school attendance. To estimate the impact of the program on the equilibrium wage, Attanasio, Meghir, and Santiago (2011) include a dummy variable in the wage function indicating whether a village was in the treatment sample. Given randomization of villages into treatment and control groups, the difference in their mean (offer) wages must be due to the equilibrium effects of the program. Evaluating the effect of the program involves including both the subsidy for school attendance and a concomitant increase in the child wage. However, the equilibrium effect that is estimated in this way is the effect only for the actual subsidy schedule implemented in the PROGRESA program.

To perform counterfactual exercises that change the subsidy schedule, one also needs to estimate how the child wage changes with the quantity of child labor; that is, one needs an estimate of the village-level demand for child labor. Attanasio, Meghir, and Santiago (2011) estimate the relative child-to-adult labor demand function using the random assignment of villages to treatment and control groups to generate exogenous variation in child labor supply.[69] A full-information estimation procedure, not pursued by Attanasio, Meghir, and Santiago (2011), would explicitly solve for the equilibrium wage in the estimation of the school attendance decision model, that is, aggregate the school attendance decisions that solve the optimization problem of each household in the village to calculate the supply of child labor and then find the wage that equates the supply and demand for child labor.[70]

The Attanasio, Meghir, and Santiago (2011) model is dynamic given that school attainment, that is, the accumulation of past school attendance decisions, directly affects alternative-specific utilities and also because the subsidy amount varies with grade level. The school/work decision is made at each age from 6 to 17, at which time there is a terminal payoff that depends on the number of years of schooling completed. Child wages are observed only for those who work. The (log) child wage offer received in each period depends on the child's education, age, the village-level minimum wage, and, as discussed above, the village's treatment status. Attanasio, Meghir, and Santiago (2011) also allow for failure; the probability of passing a grade given attendance depends on the grade level and the child's age.[71] The difference in independent extreme value preference errors is logistic; fixing the scale parameter of the logistic, the parameters represent normalized

effects on utilities. The alternative specific value functions, taking into account the dynamics, are specified similarly to those already discussed.

Attanasio, Meghir, and Santiago (2011) follow a two-stage procedure in estimating the household decision model. In the first stage they estimate the wage offer function, using a Heckman two-step selectivity correction and the school failure probability function.[72] In the second stage they substitute the predicted wage obtained from the first stage into the flow utility function and take the parameters of the failure probability function as given in solving the dynamic programming problem. The implicit assumption in making this predicted wage substitution is that the stochastic component of the wage offer is observed only after the school/work decision has been made.[73] The extreme value error assumption implies that the DP problem as well as the choice probability have closed-form representations (Rust, 1987). Attanasio, Meghir, and Santiago (2011) deal with the initial conditions problem, that is, that years of schooling completed, the outcome of prior school attendance decisions, is related to the child-specific fixed factor $(\mu_i^s - \mu_i^w)$, by positing an ordered probit as governing the determination of preprogram years of schooling that also includes the child-specific fixed factors.[74] As in Todd and Wolpin (2006), this procedure can be justified as an approximation to the probability distribution of accumulated schooling that would be the outcome of solving the dynamic programming problem back to the "true" initial period (say, the date at which the firstborn child is age 6, the school entry age) and then integrating school attendance decisions up to the first observation period.

2.2.2.3 *Results*

Table 2.4 presents the estimated impacts of the PROGRESA subsidy on attendance rates for the two Todd and Wolpin papers and for the Attanasio, Meghir, and Santiago paper. The Todd and Wolpin (2006, 2008) papers are *ex ante* evaluations (out-of-sample predictions), whereas the Attanasio, Meghir, and Santiago (2011) paper is an *ex post* evaluation (within-sample prediction). The estimates in Todd and Wolpin (2006) are calculated as the difference in the predicted attendance rate of the control observations if they had been given the PROGRESA subsidy schedule (out-of-sample prediction) and the predicted attendance rate of the control observations with no subsidy (within-sample prediction). Attanasio, Meghir, and Santiago (2011) calculate the difference between

Table 2.4
The Predicted and Experimental Effect of the PROGRESA Subsidy Policy on School Attendance Rates of Children Aged 12–15

	Boys		Girls	
	Predicted	Experimental	Predicted	Experimental
Todd and Wolpin (2008): S-NP[1]				
Single child	0.056	0.033	0.060	0.090
Multiple children	0.059		0.070	
Todd and Wolpin (2006): S-P	0.077	0.038	0.064	0.080
Attanasio, Meghir, and Santiago: S-P	0.065[a]	0.056	na	na

[a]This number is not reported in Attanasio, Meghir, and Santiago but is calculated by the author from a figure showing the impacts by single ages and taking a simple average.

the predicted attendance rate of the treatment observations with the subsidy (within-sample prediction) and the predicted attendance rate of the treatment observations with no subsidy (equivalent, given randomization, to the within-sample prediction of the control observations). Table 2.4 also shows the actual impacts obtained from the experiment, which differ across the papers as a result of the different samples used in estimation. For Todd and Wolpin (2006) the comparison of the predicted and experimental effect constitutes an external validation, whereas for Attanasio, Meghir, and Santiago (2011) the comparison demonstrates within-sample fit.

As seen from the table, the difference between the predicted and experimental effects for boys is somewhat smaller for Attanasio, Meghir, and Santiago (2011) than for Todd and Wolpin (2008) and Todd and Wolpin (2006), both absolutely and proportionally. Attanasio, Meghir, and Santiago (2011) do not report estimates for girls, but the difference for girls in both Todd and Wolpin (2006, 2008) papers is about the same as that for boys in the Attanasio, Meghir, and Santiago (2011) paper. One would expect Attanasio, Meghir, and Santiago (2011) to do better because the experimental treatment effect is in the data they use in estimation, whereas that is not the case for Todd and Wolpin (2008) or Todd and Wolpin (2006). Indeed, the model specification in Attanasio, Meghir, and Santiago (2011), except for functional form and distributional assumptions, provides a direct estimate of the subsidy effect from the treatment and control observations.

2.2.2.4 *Alternative Policies*

Designing an optimal subsidy scheme to achieve some desired increase in schooling requires knowledge of the effects of alternative subsidy schedules on take-up rates. As noted earlier, a limitation of experiments is that they do not typically provide a reliable way of extrapolating to learn about effects of counterfactual policies. Although a small change in the subsidy schedule might be well approximated by a simple extrapolation of the experimental treatment effect, it is unclear on what basis extrapolations of more radical changes would be made.

Attanasio, Meghir, and Santiago (2011) and Todd and Wolpin (2006, 2008) simulated the effect of alternative policies. Attanasio, Meghir, and Santiago (2011) and Todd and Wolpin (2008) reported results for attendance rates. Todd and Wolpin(2006) also solved the optimization problem for each household over their entire life cycle, including a fertility decision at each point in time, and calculated the completed schooling of all of the children ever born.[75] The results reported by Todd and Wolpin (2006) of implementing alternative policies are shown in table 2.5. As the first column shows, the model predicts, based on simulations of households from the time of marriage until the lastborn child reaches age 16, that the average years of completed schooling in the absence of the program would be 6.29 for girls and 6.42 for boys and that 19.8 percent of girls and 22.8 percent of boys would have completed the ninth grade. The first counterfactual (column 2), the perfect enforcement of a compulsory school attendance law, establishes a maximal potential program impact. Given predicted failure rates (which are lower for girls), average completed schooling would be at most 8.37 years for girls and 8.29 for boys. The next column shows the predicted effects of the PROGRESA subsidy schedule. The model predicts an increase in completed schooling of about one-half year for both boys and girls, or 26.0 percent of the maximal potential increase for girls and 28.9 percent for boys. The last row of the table reports the per-family government budgetary cost of the program over the lifetime of the families in the sample, that is, from the woman's age at marriage to age 59. The model predicts that cost per family to be 26,000 (1997) pesos.

As noted, the PROGRESA subsidy schedule rewards school attendance starting at grade 3. However, attendance in grades 3–5 is almost universal, making this aspect of the program essentially a pure income transfer. Todd and Wolpin (2006) calculated that the per-family cost of the program could be held roughly constant if the subsidy in grades 3–5 were eliminated and the subsidy in grades 6–9 were increased by

Table 2.5
The Effectiveness and Cost of Alternative Education Policies: SP Estimation[a]

	Baseline	Compulsory Attendance	Original Subsidy	Modified Subsidy	Bonus for Completing Grade 9	Build Schools	Income Transfer (5,000 pesos/yr)	Subsidy + 25% Increase in Child Wage
Mean Completed Schooling								
Girls	6.29	8.37	6.83	6.97	6.50	6.39	6.41	6.75
Boys	6.42	8.29	6.96	7.07	6.58	6.55	6.53	6.79
Cost per Family	0	?	26,096	25,193	36,976	?	237,000	25,250

[a]Todd and Wolpin (2006).

about 45 percent. Column 4 shows that the cost of the grade 3–5 subsidy in terms of forgone completed schooling would be 0.14 years for girls and 0.11 years for boys. Moreover, under the modified plan, the proportion of girls completing ninth grade would increase by 3.4 percentage points, and the proportion of boys by 3.8 percentage points, although there would also be a small decline in the proportion of children who complete at least the sixth grade.

An alternative subsidy scheme would reward grade completion rather than attendance. As the next column in table 2.5 shows, the impact of a ninth-grade graduation bonus of 30,000 pesos would have a relatively small impact on average schooling, 0.21 years for boys and 0.26 years for girls. In fact, the increase in average schooling is not as large as the effect of the original subsidy even though the cost of the bonus program is about 40 percent higher.[76]

Todd and Wolpin (2006) also simulate the effect of building a grade 7–9 school in each village where its absence is indicated by setting to zero the distance of such a school from the village. As seen in table 2.5, this intervention would raise mean schooling by 0.10 years for boys and 0.13 for girls. They also simulate the impact of a pure income transfer program, one that pays 5,000 pesos per year to families without any school attendance requirement.[77] Under that policy, mean schooling would increase by a little over 0.20 years for both genders. However, that increase in schooling is only about 20 percent as large as the original attendance-based subsidy. Moreover, its cost per family is an order of magnitude larger.

Attanasio, Meghir, and Santiago (2011) performed two similar counterfactuals. As in Todd and Wolpin (2006), (1) they simulated the impact of eliminating the subsidy to primary school and redistributing the savings to increase the subsidies at later grades, and (2) they simulated the impact of building schools.[78] As did Todd and Wolpin (2006), they found the effect of the first experiment to be large. The modified program would increase school attendance rates of boys from 0.065 to 0.106, an increase of about 70 percent.[79] Attanasio, Meghir, and Santiago (2011) found a modest effect of building schools relative to the first experiment, as did Todd and Wolpin (2006). The similarity of the findings between these studies in both experiments is perhaps surprising given the very significant differences between the model structures and estimation samples.

As discussed, Attanasio, Meghir, and Santiago (2011) accounted for general equilibrium effects on the child wage. They found that the child

wage increased 6 percent due to the withdrawal of child labor induced by the PROGRESA school attendance subsidy. Attanasio, Meghir, and Santiago (2011) report the effect of taking into account the increased child wage as only minimally reducing the impact of the program on school attendance. Todd and Wolpin (2006), having estimated the treatment effect assuming no general equilibrium effects, performed a counterfactual experiment, as shown in the last column of table 2.5, in which the wage was increased by 25 percent, four times the increase estimated by Attanasio, Meghir, and Santiago (2011). Todd and Wolpin (2006) find the increase in completed school would be 85 percent of the partial equilibrium effect of the original subsidy for girls and 69 percent for boys.[80]

Todd and Wolpin (2006, 2008) and Attanasio, Meghir, and Santiago (2011) performed the identical counterfactual policy experiment of doubling the subsidy at all grades.[81] The first column of table 2.6 repeats the results from table 2.4 showing the predicted effect of the actual subsidy on the attendance rate of 12- to 15-year-olds (by gender). As seen in column 2 of the table, three of the predictions are remarkably close, particularly given the different behavioral models, estimation samples, and estimation methods: the single-child matching estimator in Todd and Wolpin (2008) implies that doubling the subsidy would increase the attendance rate for boys by a factor of 2.08, the estimate from the DCDP model of Todd and Wolpin (2006) implies an increase by a factor of 2.06, and the DCDP model of Attanasio, Meghir, and Santiago (2011) an increase by a factor of 1.87. The multiple-child

Table 2.6
The Predicted Effect of Doubling the PROGRESA Subsidy on School Attendance Rates of Children Aged 12–15

	Boys		Girls	
	1 × Subsidy	2 × Subsidy	1 × Subsidy	2 × Subsidy
Todd and Wolpin (2008): S-NP[1]				
Single child	0.056	0.116	0.060	0.141
Multiple children	0.059	0.078	0.070	0.089
Todd and Wolpin (2006): S-P	0.077	0.159	0.064	0.146
Attanasio, Meghir, and Santiago: S-P	0.070	0.131	na	na

matching estimate of Todd and Wolpin (2008), on the other hand, implies an increase by a factor only of 1.32. Of course, the closeness of the three estimates does not by itself imply that they provide a more accurate prediction.[82]

In assessing the value of the DCDP structural approach, it is useful to recall the finding in both papers about the effect on schooling of a budget-neutral shift in resources toward the higher grades. That such a shift would increase schooling overall must be true from the fact that the attendance up to grade 5 is essentially universal without the subsidy. However, it is not possible to determine the impact quantitatively from the experiment alone. A policy maker with limited resources could not make an informed decision about whether to continue the subsidy to the lower grades, given its redistributive function, without knowing its quantitative tradeoff with forgone schooling. The structural approach permits a quantitative cost–benefit analysis of alternative programs without having to conduct costly field experiments.

2.3 Is the Use of Holdout Samples for Validation and/or Model Selection Justified?

Should one have a stronger belief in the predictions of the counterfactual experiments from the Todd and Wolpin (2006, 2008) studies as opposed to the Attanasio, Meghir, and Santiago (2011) study because the Todd and Wolpin studies used a holdout sample for external validation? Additionally, how should a policy maker come up with an estimate of the effect of doubling the subsidy based on the four estimates? The first question can be thought of as one of model validation and the second one of model selection. Schorfheide and Wolpin (2011) present a formal framework within which these questions can be rigorously addressed.

Let me first, however, provide some background discussion.[83] One goal of empirical research is to provide evidence on the validity of decision-theoretic models that describe the behavior of economic agents. There are two approaches to this endeavor that stem from different epistemological perspectives. The first stems from a view that there exists a "true" decision-theoretic model from which observed data are generated. This leads naturally to a model validation strategy based on testing the validity of the behavioral implications (and possibly the assumptions) of the model and/or testing the fit of the model

to the data. A model is not deemed invalid if it is not rejected in such tests according to some statistical criterion. Rejected models are deemed invalid and discarded, although they may serve as building blocks for more refined models.

The second approach stems from a pragmatic view in which it is assumed that all models are necessarily simplifications of actual decision-making behavior. Hypothesis testing as a means of model validation or selection is eschewed because, given enough data, all models would be rejected as true models. There is no true decision-theoretic model—only models that perform better or worse than other models in addressing particular questions. Models should be chosen that are "best" for some specific purpose, and alternative models may be valid for different purposes.

Decision-theoretic models are typically designed and estimated with the goal of predicting the impact on economic agents of changes in the economic environment. Thus, one criterion for model validation/selection that fits within the *pragmatic view* is to examine a model's predictive (out-of-sample) accuracy, that is, how successful the model is at predicting outcomes of interest within the particular context for which the model was designed. In contrast, in the *absolutist view*, a model would be considered useful for prediction only if it were not rejected, despite the fact that nonrejection does not necessarily imply that predicted effects will be close to actual effects. Nor will nonrejected models necessarily outperform rejected models in terms of their (context-specific) predictive accuracy.

A particularly challenging use of decision-theoretic models is to forecast the impact of large changes in the environment. Often, these changes are related to policy interventions that are outside of the scope of current policies, such as the PROGRESA program. A good in-sample fit or nonrejection of a model's implication is unlikely, in itself, to give us much confidence in its forecasting ability in such contexts. This problem arises because of the common practice of using the same data both for estimation and for model development, that is, the ubiquitous and unavoidable practice of data-mining.

The idea for using a holdout sample, as in Todd and Wolpin (2006, 2008), is that if a model can provide a good forecast for a holdout sample that faces a policy regime well outside the support of the data (and that is not used in model formulation), then we should gain confidence that it can provide a good forecast of impacts of other policy changes along the same dimension (such as the doubling of the PROGRESA subsidy) or perhaps even other dimensions (such as a graduation bonus rather than an attendance subsidy).[84]

Among the earliest uses of a randomized social experiment for the purpose of model validation and selection is that by Wise (1985). Wise exploited a housing subsidy experiment to evaluate a model of housing demand. In the experiment, families that met an income eligibility criterion were randomly assigned to control and treatment groups. The latter were offered a rent subsidy. The model was estimated using only control group data and was used to forecast the impact of the program on the treatment group. The forecast was compared to its actual impact. In addition to Todd and Wolpin (2006, 2008), recent examples of this approach to validation, within a randomized controlled experimental design, include Bourgignon, Ferreira, and Leite (2003), Lise, Seitz, and Smith (2003), and Duflo, Rema, and Ryan (2012).

McFadden et al. (1977) provides an early example in which a large regime shift in a nonexperimental setting is exploited. McFadden estimated a random utility model (RUM) of travel demand before the introduction of the Bay Area Rapid Transit (BART) system, obtained a forecast of usage, and then compared the forecast to actual usage after BART's introduction. McFadden's model validation treats pre-BART observations as the estimation sample and post-BART observations as the validation sample.

McFadden did not purposefully hold out the post-BART data; the point of the analysis was to forecast usage prior to BART having been built. A number of later papers have mimicked that validation design by purposefully holding out a part of the sample. Lumsdaine, Stock, and Wise (1992) estimated a model of retirement behavior of workers in a single firm who were observed before and after the introduction of a temporary 1-year pension window. They estimated several models on data before the window was introduced and compared the forecast of the impact of the pension window on retirement based on each estimated model to the actual impact as a means of model validation and selection. Keane and Moffitt (1998) estimated a model of labor supply and welfare program participation using data after federal legislation (Omnibus Budget Reconciliation Act , 1981) that significantly changed the program rules. They used the model to predict behavior prior to that policy change. Keane (1995) used the same model to predict the impact of planned expansions of the Earned Income Tax Credit in 1994–1996. Keane and Wolpin (2007) estimated a DCDP model of welfare take-up (along with choices about schooling, fertility,

marriage, and work) using a set of five states, with medium to high welfare generosity, and validated the model based on the forecasts of those choices in a low-welfare state. Cohen-Goldner and Eckstein (2010) estimate a model of the assimilation of female immigrants to Israel from the former Soviet Union using data from the first 5 years of Israeli residence and validate the model using 10 years of data from samples not used in estimation. Kaboski and Townsend (2011) structurally estimated a model of credit-constrained households deciding on consumption, indivisible investment, and savings. They estimated the model using data collected prior to the introduction of a large-scale government microfinance program, the Thai Million Baht Village Fund Program, and then validated the model using postprogram data.

The common and essential element in all of these validation exercises, regardless of whether they are based on randomized experiments or whether the researcher has chosen a nonrandom holdout sample, is the existence of some form of regime change radical enough to provide a degree of distance between the estimation and validation samples.[85] The more different the regimes, the less likely are forecasted and actual behavior in the validation sample to be close purely by chance. Although this latter statement is true, it does not in itself justify the loss of observations in the estimation sample.

Because the work reported by Schorfheide and Wolpin (2011) is a work in progress, I provide only a very brief discussion. Schorfheide and Wolpin (2011) consider the following problem of a policy maker who has run an experiment, like PROGRESA. The policy maker recognizes that the experiment provides an estimate only of the impact of the specific experimental policy, for example, the impact of the original PROGRESA subsidy schedule. But, the policy maker would like to know what would have happened had a different subsidy level been chosen, for example, had it been doubled, or if the gradient of the subsidy with grade level had been different. The policy maker also realizes that there are researchers who have different models that can be used to extrapolate effects to these alternative policies. For concreteness, the policy maker is going to decide whether or not to double the size of the subsidy when the program is implemented.

The policy maker would like to be able to assess the relative validity of different models in deciding which estimate to use or how to combine the estimates. From a Bayesian perspective, the best the policy maker can do is to give researchers all of the data from the experiment, from both the treatment and controls, and evaluate the models on the basis

of the marginal likelihood of attendance. The policy maker can then model average to calculate the effect of doubling the subsidy. The policy maker, however, recognizes that the researchers may data mine and not truthfully report their marginal likelihoods.

The policy maker has the option of holding out part of the data, for example, giving researchers only the control group data, as opposed to giving the researcher all of the data from both groups. If the policy maker holds out part of the sample, then the policy maker would request from researchers the predicted likelihood for the holdout sample and the predicted density for the treatment effect. Note that the policy maker knows the actual distribution of attendance in the holdout sample and the average treatment effect based on all of the data. Schorfheide and Wolpin (2011) represent data mining by the researcher who is given data from both groups as data-based modification of the researcher's prior distribution of the subsidy effect. In this way, the researcher is able to increase the marginal likelihood that is reported to the policy maker. If the policy maker gives the researcher only the control data, then the researcher cannot data mine. Schorfheide and Wolpin (2011) show by simulation that there are circumstances in which the policy maker with a quadratic loss function would be better off by providing only data from the control sample.