Journal of Economic Literature 2023, 61(1), 41–85
https://doi.org/10.1257/jel.20211652

The Best of Both Worlds: Combining Randomized Controlled Trials with Structural Modeling†

Petra E. Todd and Kenneth I. Wolpin*

There is a long-standing debate about the extent to which economic theory should inform econometric modeling and estimation. This debate is particularly evident in the program/policy evaluation literature, where reduced-form (experimental or quasi-experimental) and structural modeling approaches are often viewed as rival methodologies. Reduced-form proponents criticize the assumptions invoked in structural applications. Structural modeling advocates point to the limitations of reduced-form approaches in not being able to inform about program impacts prior to implementation or about the costs and benefits of program designs that deviate from the one that was implemented. In this paper, we argue that there is a new emerging view of a natural synergy between these two approaches, that they can be melded to exploit the advantages and ameliorate the disadvantages of each. We provide examples of how data from randomized controlled trials (RCTs), the exemplar of reduced form practitioners, can be used to enhance the credibility of structural estimation. We also illustrate how the structural approach complements experimental analyses by enabling evaluation of counterfactual policies/programs. Lastly, we survey many recent studies that combine these methodologies in various ways across different subfields within economics. (*JEL* C21, C52, C53, H24, I38, J13, R38)

1. *Introduction*

The use of modern econometrics and computational methods in the practice of empirical economics research has stimulated much debate. The history of this debate, spanning many decades, is exemplified by the titles of the following influential books: *Measurement Without Theory* (Koopmans 1947), *Specification Searches: Ad Hoc Inference with Nonexperimental Data* (Leamer 1978), and *Mostly Harmless Econometrics: An Empiricist's Companion* (Angrist and Pischke 2009). More recently,

*Todd: Edmund J. and Louise W. Kahn Term Professor of Economics at the University of Pennsylvania. Wolpin: Emeritus Professor of Economics at the University of Pennsylvania. This paper was presented at the Society of Labor Economics 2020 meetings as the SOLE Fellows keynote lecture. We thank Robert Moffitt and Steven Durlauf and three anonymous referees for helpful suggestions.

† Go to https://doi.org/10.1257/jel.20211652 to visit the article page and view author disclosure statement(s).

attention has focused on the choice of empirical methodologies for conducting research in policy/program evaluation. The distinguishing feature of alternative evaluation approaches is the extent to which economic theory informs econometric modeling and estimation.

Popular terminology identifies one evaluation approach as "reduced form." A common aim of that approach is to estimate the impact of existing programs or policies. The reduced-form approach often invokes the notion of an "experiment" in that there is an identifiable group that is subject to the program or policy, a treatment group, and another group that is not, a comparison or control group. Reduced-form analyses that are based on an explicit randomization, a randomized controlled trial (RCT), or on a (so-called) natural experiment are deemed to be experimental. Analyses not based on an explicit or natural randomization are sometimes called quasi-experimental (for example, the use of difference-in-difference, matching or regression discontinuity methodologies). A common aim of reduced-form evaluations is to estimate the impacts of existing programs or policies. A second evaluation approach is popularly termed "structural," which generally consists of a fully specified behavioral model, usually, though not necessarily, parametric. The structural approach is often used to evaluate existing policies and to perform counterfactual program/policy experiments, such as the evaluation of new hypothetical policies.[1]

Reduced-form and structural approaches have long been considered to be rival methodologies for conducting empirical economics research (e.g., Heckman 2001, Angrist and Pishke 2010). Proponents of the reduced-form approach criticize the assumptions invoked in structural applications, whereas proponents of the structural approach point to limitations of reduced-form analyses, such as not being able to inform about program impacts prior to implementation or about the costs and benefits of program designs that deviate from an existing program. In this paper, we argue the merits of an emerging view, that there is a natural synergy between experimental and structural approaches. We review a new literature, which evolved over the last two decades, that combines these two approaches to exploit the advantages and ameliorate the disadvantages of each.

Pioneering studies by Wise (1985) and Lalonde (1986), precursors to the more recent literature, were among the first to exploit synergies between experiments and structural estimation. Wise (1985) exploited a housing subsidy experiment to evaluate a housing demand model. In the experiment, families that met an income eligibility criteria were randomly assigned to either a control or treatment group, where the latter was offered a rent subsidy. Wise estimated the housing demand model using only control group data, used the model to forecast the program impact on the treatment group, and compared the forecast to the impact measured by the RCT. Wise's approach to combining RCTs and structural estimation to analyze the performance of structural models was not readily pursued by other researchers until several decades later, beginning, as far as we know, with Todd and Wolpin (2006).

[1] A critical feature of the structural approach in performing ex ante evaluation, that is, the evaluation of a program that has not been implemented or is an untried modification of an existing program, is structural invariance (Marschak 1953, Lucas 1976). Ex post evaluation, that is, evaluation of an existing program, makes use of actual policy variation, for example, observations on different individuals with and without the program or observations on the same individuals before and after a program is implemented. In either type of evaluation, the structural approach fully specifies how the behavioral model is altered due to the program. Structural invariance is not relevant for reduced-form ex post analysis, which generally does not specify the mechanisms through which the program affects outcomes.

Lalonde (1986) used an RCT, the National Supported Work (NSW) Demonstration training program, to test the validity of alternative nonexperimental estimators of program impacts. The estimators he considered were ex post and made use of both the treatment group and the comparison group data. He found that different nonexperimental estimators yielded different impact estimates and, furthermore, that the estimates deviated substantially from the experimental benchmarks. His research spawned an immediate literature further examining the performance of alternative nonexperimental estimators in comparison to RCT estimates and devising tests to choose among them. For example, Heckman and Hotz (1989) developed preprogram exogeneity tests that were useful in narrowing the range of nonexperimental estimates, although a wide range remained after applying these tests.[2]

The perceived failure of nonexperimental methods to reproduce experimental results added to a prior literature critiquing the value of tightly connecting economic theory to estimation.[3] Taken together, this cumulative body of work helped spur a movement that rejected the use of the structural approach, based on formal economic modeling, in favor of a reduced-form, purely statistical, approach.[4] Angrist and Pishke (2010) declared that a "credibility revolution" took place with researchers increasingly relying on experimental and quasi-experimental research designs.

Implementing a reduced-form ex post evaluation requires data on a treated group and on an untreated comparison group. Given the program evaluation goal, the main threat to validity in comparing the outcomes of the two groups is nonrandom treatment selection. To obtain a reliable treatment effect estimate requires either random assignment to treatment, an assumption that selection is on observables, functional form assumptions on unobservables, or some exogenous element in the assignment rule, such as a lottery that provides the basis for an instrument.[5] If done well, an RCT provides an unbiased estimate of average treatment effect under minimal assumptions. RCTs can also be used to examine treatment impact heterogeneity in a straightforward way when the sample sizes are sufficient to permit subgroup analyses. However, as noted in prior research, RCTs also have some significant limitations.[6]

In the context of this essay, the most relevant limitation of RCTs is their limited scope. RCTs are often costly, which makes it infeasible to extensively vary the treatment design or the length of treatment exposure within the experiment. Most often there is a single treatment as, for example, in the Mexican conditional cash transfer program, PROGRESA, studied by Todd and Wolpin (2006) and by Attanasio, Meghir, and Santiago (2012), and in the Indian teacher incentive program studied by Duflo, Hanna, and Ryan (2012). Researchers may be interested in the potential impacts and costs of a range of hypothetical programs with different design parameters, particularly if interest centers

[2] Heckman, Ichimura, and Todd (1997) argued that one reason that the estimators Lalonde (1986) considered did not perform well was that his data were not rich enough and that the econometric models perform better with better data.

[3] Results based on the estimation of demand systems, at least as far back as Stone (1954), generated a large literature questioning the empirical value of the neoclassical model of demand (see, for example, Blaug 1980).

[4] The reduced-form approach is sometimes referred to as "causal modeling," even though it eschews the modeling of mechanisms.

[5] Heckman and Urzua (2010) provide a critical assessment of the role of instrumental variables in answering relevant economic questions.

[6] Leamer (1983) provides an early discussion of the interpretation of experimental results. Deaton (2010) critically reviews the role of field experiments in development economics. See Imbens (2010) for a response to both Deaton (2010) and Heckman and Urzua (2010).

around designing a program that achieves some optimality criteria for a given budget. RCTs provide information on the particular design that was implemented and are typically uninformative about the potential costs and benefits of alternative program designs.

In the structural approach, the researcher specifies and estimates a formal economic behavioral model.[7] The model structures that researchers use vary according to the policy issue being addressed and include static or dynamic single-agent or game-theoretic models as well as partial equilibrium or general equilibrium frameworks. A key limitation of the structural approach is that the estimation almost always relies on additional atheoretic assumptions about functional forms and error distributions, usually chosen partly for computational convenience. More fundamentally, researchers may disagree on the appropriate behavioral framework.[8] Perhaps the most vexing problem in empirical research is that of model validation and selection.

In the program evaluation context, and depending on the model specification, structural methods can be used for (i) simulating program impacts, costs and take-up rates under alternative program designs; (ii) analyzing the behavioral mechanisms that generate observed outcomes and program impacts and quantifying welfare effects; (iii) analyzing program impacts over a time horizon that exceeds the length of time observed in the data; (iv) analyzing program impacts in the presence of spillover or general equilibrium effects; and (v) analyzing the effect of extending the program to different populations. In some cases, the structural method can also be used for ex ante evaluation purposes, that is, to predict the effects of a program intervention prior to its implementation, making it possible to study the potential impacts and costs of alternative program designs prior to implementing them.

This paper builds on a previous *JEL* survey by Heckman (2010) that described ways of "building bridges" and finding a "middle ground" between structural modeling and reduced-form program evaluation approaches. In that survey, Heckman makes explicit the economics implicit in local average treatment effect (LATE) evaluation approaches, and he proposes methods for moving beyond LATE to identify and estimate parameters that he argues are of greater policy relevance. Drawing on a theorem of Vytlacil (2002) that shows that the LATE model of Imbens and Angrist (1994) is equivalent to a nonparametric version of the generalized Roy model, Heckman (2010) provides an economic interpretation of LATE within the Roy model framework. He surveys methods developed in Heckman and Vytlacil (2005), Heckman, Urzua, and Vytlacil (2006), Cunha, Heckman, and Navarro (2007), and Carniero, Hansen, and Heckman (2003) for generalizing and extending LATE analysis for two-outcome and multiple-outcome models, including ordered and unordered choice models, and he introduces policy relevant treatment effects (PRTE). Heckman (2010) emphasizes the value of placing the policy questions foremost and asking how the questions can be answered with statistics, rather than focusing on what parameters can be easily obtained with statistics and then asking if they happen to be policy relevant.

More recently, Mogstad, Santos, and Torgovitsky (2018) develop methods that build on the marginal treatment effect (MTE) and PRTE estimators analyzed in the papers by James Heckman and his

[7] The theoretical basis for these models spans both neoclassical and behavioral economics. The surveys by Keane, Todd, and Wolpin (2011), primarily of the former, and DellaVigna (2018), of the latter, provide a number of examples.

[8] An example would be the choice of a unitary, collective or noncooperative model of household decision-making.

coauthors. Using the concept of a "marginal treatment response" (MTR), they show how to extract information from a class of "instrumental variables (IV)-like" estimands to construct nonparametric bounds on the average causal effect of certain kinds of hypothetical policy changes.[9] The methods they develop use multiple instruments to enable extrapolation of average treatment effects of compliers to different subpopulations of interest. Usually, the estimators deliver bounds on the parameter of interest, although in some cases, depending on the variation in treatment assignment induced by the instruments, they obtain point identification.

This paper takes the policy questions at the stages of designing, implementing, and refining a program to be the central focus and shows how behavioral models can be used to address such questions. The models we describe typically specify, in greater detail than in Heckman's papers or in the Mogstad, Santos, and Torgovitsky (2018) paper, the theoretical mechanisms that determine outcomes and choices as well as program components. Imposing additional structure and functional form assumptions carries a risk of model misspecification, but it also provides the framework needed to carry out ex ante policy evaluation, to analyze a wide array of changes to the design of program components, and to accommodate possible general equilibrium effects. We survey a variety of approaches developed in the recent literature for combining RCT data with structural modeling to increase the credibility of inference from such models and to significantly expand the scope of questions that researchers can address. We illustrate various approaches by examining recent applications spanning a number of subfields within economics.[10]

There are two ways that RCT data can be used to enhance the credibility of structural methods. The first is for purposes of model validation and selection, using either the treatment group or the control group as a "holdout" sample in performing out-of-sample model fit tests. Such a strategy mitigates the impact of data mining that is inherent in the formulation of structurally estimated empirical models and that limits the applicability of standard model selection criteria.[11] When model parameter estimation is feasible without treatment variation (see examples below), the estimated model can be used to forecast the choices and outcomes of the holdout sample and the forecasts compared to the actual holdout sample data. If the model forecasts are "sufficiently" accurate, then the model is deemed to fit well and to be potentially useful for other purposes, such as analyzing the effect of varying parameters of the program's design. If the model does not generate accurate forecasts, then the researcher knows that the problem lies with the model, because the randomization ensures that the distribution of unobservables is the same in the treatment and control samples. A second way that researchers can use RCT data is to base estimation on both the treatment and control group. In this case, variation induced by the treatment provides an additional, and sometimes necessary, source of variation for identifying and estimating model parameters and improving

[9] For example, some of the policies they consider are assumed to affect the decision rule to participate in a treatment, but to not affect outcomes directly, so they represent exclusion restrictions in a generalized Roy model framework.

[10] A parallel literature exploits (presumed exogenous) policy regime shifts in a manner similar to RCTs. Although our main focus is on RCTs, we present examples from that literature as well.

[11] See Keane and Wolpin (2007) and Schorfheide and Wolpin (2012, 2016).

precision.[12] These two approaches to using the RCT data can be combined, that is, researchers can first use either the control group or treatment group data as a holdout sample and then afterwards reestimate the model using both groups.

A requirement for combining these approaches is that the experimental data go beyond measurement of treatment status and outcomes. Successful empirical implementation of behavioral models requires that the key variables governing decision-making, as described by the model, be measured. For example, as part of the PROGRESA experiment in Mexico, the government collected extensive survey data from the families in both treatment and control villages, which allowed researchers to implement reduced-form modeling strategies (including RCT, regression discontinuity, and matching estimators) as well as to specify and structurally estimate rich models of family behavior that allow for counterfactual program analysis.

This paper develops as follows. Section 2 describes two earlier strands of literature that laid the groundwork for forecasting policy effects using behavioral models and for evaluating the models' performance against experimental benchmarks. Section 3 illustrates how and when structural models can be used for ex ante evaluation, discusses both nonparametric and parametric approaches, and considers the value of incorporating RCT data. Section 4 describes alternative approaches to assessing model validity. Section 5 surveys many recent studies across different subfields within economics that combine RCT/quasi-experimental and structural modeling approaches in different ways. Section 6 focuses on a smaller set of recent papers that develop models to account for spillover effects or general equilibrium effects in evaluating policy effects. Section 7 concludes.

2. *Early Related Literature*

2.1 *Studies of the Reliability of Models to Forecast Decision-Making*

The problem of forecasting the effects of hypothetical social programs is part of the more general problem of studying the effects of policy changes prior to their implementation that was described by Marschak (1953) as one of the most challenging issues facing empirical economists.[13] In practice, in the early discrete choice literature, researchers used random utility models (RUMs) to predict the demand for a new good prior to its being introduced into the choice set.[14] Both theoretical and empirical criteria were applied to evaluate model performance. Empirically, a model's performance could sometimes be assessed by comparing the model's predictions about demands for good with the ex post realized demand. In one of the earliest applications of this idea, McFadden et al. (1977) used a RUM to forecast the demand for the San Francisco BART subway system prior to its opening, and then checked the forecast's accuracy against the actual subway demand data. A later study by Lumsdaine, Stock, and Wise (1992) studied the performance of alternative models at forecasting the impact of a pension bonus program on older workers' retirement. The authors first estimated the models using data gathered prior to the bonus program and then compared the models' forecasts to actual data on workers' departures.

[12] In some of the applications described below, using the data variation induced by the randomized treatment permits identification of some model parameters without having to make additional exclusion restrictions.

[13] See Heckman (2001).

[14] Much of the initial empirical research was aimed at predicting the demand for transportation modes.

The earliest examples of empirical studies comparing treatment effect estimates based on structural models to those obtained from randomized experiments were studies related to a set of negative income tax (NIT) experiments conducted in the 1970s, probably the most heavily studied randomized experiments in economics. For example, Moffitt (1979) used a labor supply model to forecast the effects of the Gary Negative Income Tax Experiment, which provided wage subsidies and income guarantees to low-income people. Burtless (1987) provides a summary of many of comparisons of the model-based estimates and RCT estimates, concluding that nonexperimental estimates of the responsiveness of hours worked to the tax rate are somewhat higher than those obtained from the experiments. Because of design and other issues, it was not presumed that the estimates from the experiment were necessarily superior to the nonexperimental estimates.

The literature is vast, but to the best of our knowledge, there was no holdout sample validation exercise conducted using the NIT. As previously noted, as far as we are aware, the first use of a holdout sample in the context of a randomized experiment was Wise (1985).

2.2 *Studies of the Reliability of Nonexperimental Evaluation Estimators*

As previously noted, there has been a long-standing debate in the literature over whether social programs can be reliably evaluated without a randomized experiment. Several of the early papers were in the context of evaluating job training programs. Lalonde's (1986) influential paper compared the performance of some standard econometric estimators against RCT benchmarks using data from the NSW experiment.[15] The evaluation estimators he considered included cross-section, difference-in-difference, and control function regression estimators applied to treatment data from NSW and comparison group samples drawn from the Panel Survey of Income Dynamics (PSID) and from the Current Population Survey (CPS). He found that the impact estimates differed across estimators and that the resulting range of estimates was too wide to be useful. Heckman and Hotz (1989) developed preprogram exogeneity tests that could be applied to rule out particular estimators. The approach they suggested is to estimate treatment effects using preprogram data, when the program effects are known to be zero. Deviation from zero is taken as indicative of the estimator being biased. These tests reduced the range of the nonexperimental point estimates, although it was still substantial.

Dehejia and Wahba (1999, 2002) also analyzed the NSW data, applying a class of estimators based on propensity-score matching.[16] They found small biases and argued that matching estimators are more reliable than traditional econometric methods in reproducing the RCT results. However, Smith and Todd (2005), in a reanalysis of the data, found the Dehejia and Wahba's (1999, 2002) results to be highly sensitive to their sample selection criteria.[17]

Heckman, Ichimura, and Todd (1997) and Heckman et al. (1998) applied matching estimators to data from the Job Training Partnership Act (JTPA) experiment. They

[15] The NSW program provided job training to unemployed, urban disadvantaged populations.

[16] These estimators were introduced in the statistics literature by Rosenbaum and Rubin (1983). Traditional propensity-score matching methods pair each program participant with a single nonparticipant, where pairs are chosen based on the degree of similarity in the estimated probabilities of participating in the program (the propensity scores).

[17] Most estimators, including the standard regression estimators considered by Lalonde (1986), exhibit small biases in the data subsamples used for their analysis.

show that data quality is crucial to the performance of the estimator. The estimators were found to perform well in replicating RCT results only when they were applied to comparison group data satisfying the following criteria: (i) the same data sources (i.e., the same surveys or the same type of administrative data or both) are used for participants and nonparticipants, (ii) participants and nonparticipants reside in the same local labor markets, and (iii) the data contain a rich set of variables relevant to modeling the program participation decision. If the comparison group data fails to satisfy these criteria, the performance of the nonexperimental estimators in replicating experimental benchmarks diminishes greatly.

Glewwe et al. (2004) estimated effects of introducing flip-charts in schools using both an RCT and a difference-in-difference approach. The RCT indicated the experimental treatment effect to be essentially zero in magnitude and precisely estimated, but the difference-in-difference estimator did not replicate the RCT results.

A recent study by Griffen and Todd (2017) compared experimental Head Start Impact Study treatment effect estimates to nonexperimental estimates obtained using comparison group data from the Early Childhood Longitudinal Study, Birth Cohort (ECLS-B). They applied both conventional regression evaluation estimators and matching estimators. Some of the estimators closely reproduced the experimental results, particularly for the child test score outcomes. The difference-in-differences matching estimator exhibited the best overall performance in terms of low bias values and in capturing the pattern of statistically significant treatment effects.

In summary, the question of whether nonexperimental estimators offer a viable alternative to RCTs is still a matter of some debate. However, much evidence has been accumulated to provide guidance as to when a nonexperimental approach is likely to be successful. Having high-quality survey data and a comparison group that is highly comparable to the treated group are important to any reliable estimation strategy. The goal of this literature on nonexperimental estimators has largely been to estimate the effect of existing programs. The frameworks developed do not specify the mechanisms through which the treatment effect occurs and, in most cases, are not suitable for studying the effects of modifying a program's design.

3. *Example of How to Use the Structural Approach to Perform an ex ante Evaluation and to Analyze the Effects of Alternative Policy Designs*

In this section, we illustrate with an example of a welfare program how the structural approach can be used for purposes of ex ante evaluation and for studying effects of alternative policy designs. First, we describe a nonparametric structural approach that is feasible when a program has a particular representation in terms of the pre-program budget constraint. Second, we describe a parametric structural approach that is more broadly applicable in terms of the varieties of programs that can be analyzed. Third, we discuss how RCT data can be incorporated and the circumstances under which one of the experimental groups can be used as a holdout sample.

3.1 *A Simple Model of Welfare Participation:*

3.1.1 *Nonparametric ex ante Evaluation*

We consider two states of the world, the current state where there is no welfare program and a hypothetical state with a welfare program. In the hypothetical state, there is a welfare benefit, $b(y_i, n_i) \geq 0$, offered to unmarried women who have children and do not work outside of the home; the benefit level depends on the woman's (denoted by i) nonearned income y_i and on the number of children n_i. In either state of the world, the woman decides whether to work or not. If she works $L_i = 0$, and if not $L_i = 1$. The woman's utility function, which is assumed not to depend on the state of the world (although see below), is given by

$$U_i = U(C_i, L_i; \epsilon_i) \tag{1}$$

where C_i is woman i's consumption and ϵ_i shifts her marginal utility of leisure relative to consumption. In the current state, a woman faces the budget constraint

$$C_i = y_i + w_i(1 - L_i), \tag{2}$$

where w_i is the woman's wage if she chooses to work. In the hypothetical state, the budget constraint reflects the additional potential income from the welfare program, namely

$$C_i = y_i + w_i(1 - L_i) + b(y_i, n_i) L_i. \tag{3}$$

A woman works if $U_i(L_i = 0 | y_i, w_i, b(y_i, n_i), \epsilon_i) \geq U_i(L_i = 1 | y_i, , w_i, b(y_i, n_i), \epsilon_i)$, where $b(y_i, n_i) = 0$ in the current state.[18] If the program is offered, the take-up rate depends on the number of eligible women (for whom $b(y_i, n_i) > 0$) who choose not to work. The model implies that a woman who chooses not to work without the welfare program, and who is eligible for the program, is always better off choosing to take up welfare. We later consider the consequences for ex ante evaluation in an augmented model where there may be a stigma effect of taking welfare.

The basis for a nonparametric estimator stems from the simple insight that the budget constraint in the hypothetical state can be rewritten as

$$\begin{aligned} C_i &= \big(y_i + b(y_i, n_i)\big) \\ &\quad + \big(w_i - b(y_i, n_i)\big)(1 - L_i) \\ &= \tilde{y}_i + \tilde{w}_i(1 - L_i). \end{aligned}$$

Comparing equation (2) to (4), it can be seen that the form of the budget constraint is identical for both states of the world, with and without the welfare program. Under the assumption that the unobservable preference shifter (ϵ_i) is statistically independent of all observables, the implication of this observation is that given data in the no-welfare state, the effect of the hypothetical program can be estimated by comparing the employment status of women who have n_i children, nonearned income y_i, and wage offer w_i to women also with n_i children, but with nonearned income $\tilde{y}_i = y_i + b(y_i, n_i)$ and wage offer $\tilde{w}_i = w_i - b(y_i, n_i)$.

Todd and Wolpin (2008) develop a matching estimator that can be used to recover the effect of the program for the situation where the program can be represented as a parameterization of the existing budget constraint. Letting $H_i(y_i, w_i, b(y_i, n_i), \epsilon_i) = 1 - L_i(y_i, , w_i, b(y_i, n_i), \epsilon_i)$, the matching estimator of the policy impact on the employment rate, based on a sample of J women, is

$$\begin{aligned} \tilde{\Delta} &= \frac{1}{n} \sum_{\substack{j=1 \\ j,i \in S_P}}^{J} \tilde{E}\big[H_i | y_i = y_j + b(y_j, n_j), \\ &\qquad w_i = w_j - b(y_j, n_j)\big] - \big[H_j(y_j, w_j, n_j)\big], \end{aligned} \tag{5}$$

where S_p is the region of overlapping support.[19] For each woman, $j = 1, \ldots, J$, in the

[18] Todd and Wolpin (2008) and Wolpin (2013) consider a setting where the choice is continuous hours of work.

[19] As described in Todd and Wolpin (2008), the support restriction is needed because matches can only be found within the support of the data.

sample with observed tuple (y_j, w_j, n_j), we average the employment rate over all women with observed tuple $(y_i + b(y_i, n_i), w_i - b(y_i, n_i), n_i)$ and subtract the actual employment status of woman j.[20] The impact of the program is the average of these differences over all J women in the sample.[21]

The matching estimator can be used to analyze the impact of a menu of policies by altering the benefit schedule. Given the model, the only qualification to the estimation is that the sample needs to be large enough for the matching analysis to be credible.[22] Given a menu of alternative program designs, a policy maker can choose a design to satisfy a particular social welfare function subject to any cost constraints.

3.1.2 *Parametric ex ante Evaluation*

Extensions of the model that support fully nonparametric estimation are limited, because it is not always possible to represent programs in terms of the budget constraint in the no-program state.[23] Most researchers therefore adopt parametric models. Before considering an explicit case where nonparametric estimation is infeasible, it is useful to work through the estimation of a parametric model for the same hypothetical welfare program considered previously. The following structure establishes the conventional baseline parametric model:

$$(6) \qquad \begin{aligned} U_i &= C_i + \alpha_i L_i + \lambda C_i L_i, \\ \alpha_i &= x_i \beta + \epsilon_i, \\ C_i &= \big(y_i + b(y_i, n_i)\big) \\ &\quad + \big(w_i - b(y_i, n_i)\big)(1 - L_i), \\ w_i &= z_i \gamma + \eta_i, \end{aligned}$$

where, in addition to the terms previously defined, x_i is a vector of observed preference shifters and z_i is a vector of observed and η_i unobserved determinants of wage offers.[24] The wage function is specified to allow for the fact that only accepted wages are generally observed.[25] The employment decision is determined by a comparison of the alternative-specific utilities, $U_i(L_i = 0)$ if the women works and $U_i(L_i = 1)$ if the woman does not work:

$$(7) \quad \begin{aligned} U_i(L_i = 0) &= y_i + z_i \gamma + \eta_i, \\ U_i(L_i = 1) &= (1 + \lambda)\big(y_i + b(y_i, n_i)\big) \\ &\quad + x_i \beta + \epsilon_i. \end{aligned}$$

[20] It is actually not necessary to match on the number of children, but only on the combination of non-earned income and number of children that leads to a given welfare benefit. Matching on number of children would be necessary if fertility directly affects the work decision without welfare, for example, if the marginal utility of work depended on the number of children.

[21] The estimator can be modified to control for relevant conditioning variables by exact matching on those variables. Matches can only be performed for women whose y_j, w_j, n_j and associated $\tilde{y}_i, \tilde{w}_i, n_i$ values both lie in the support of y, w, n. Todd and Wolpin (2008) demonstrate how this matching estimator can be implemented using kernel density functions for the matching.

[22] With sufficient sample size, it would be possible to also match women in terms of demographic variables such as age, race/ethnicity/education that could affect their preferences.

[23] Wolpin (2013) discusses the viability of nonparametric ex ante evaluation under a variety of extensions of similar models, including allowing for partial observability of wages, fixed costs of work, childcare costs, kinked budget constraints, endogenous fertility, and life cycle dynamics.

[24] We adopt a linear form for the wage equation, as opposed to the more conventional log-linear form, for illustrative purposes.

[25] Todd and Wolpin (2008) and Wolpin (2013) show that a distributional assumption is required to perform an ex ante evaluation when wages are partially observed. Although the wage offer function can be estimated without distributional assumptions, the constant in the wage offer function, which is necessary for the ex ante evaluation, cannot be separately identified (see Heckman 1990, Wolpin 2013).

The latent variable function, the difference in utilities, $U_i(L_i = 0) - U_i(L_i = 1)$, is thus

$$v_i^*(x_i, w_i, \eta_i, \epsilon_i) = -\lambda\big(y_i + b(y_i, n_i)\big) + \big(z_i\gamma - b(y_i, n_i)\big) - x_i\beta + \eta_i - \epsilon_i = \xi_i^* + \xi_i$$

where $\xi_i = \eta_i - \epsilon_i$, and $\xi_i^* = -\lambda\big(y_i + b(y_i, n_i)\big) + \big(z_i\gamma_i - b(y_i, n_i)\big) - x_i\beta$.

To perform an ex ante analysis of the welfare program effects, set $b(y_i, n_i) = 0$. In that case, $\xi_i^* = -\lambda y_i + z_i\gamma_i - x_i\beta$, and the likelihood function for a sample of I women in the no-welfare state is

$$(9) \quad \pounds(\theta; x_i, z_i) = \Pi_{i=1}^{i=I} \Pr(L_i = 0, w_i | x_i, z_i, y_i)^{1-L_i} \Pr(L_i = 1 | x_i, z_{i,} y_i)^{L_i},$$

where θ is the parameter vector to be estimated, $\Pr(L_i = 0, w_i | x_i, z_i, y_i) = \Pr\big(\xi_i \geq -\xi_{it}^*(x_i, z_i, y_i) | \eta_i = w_i - z_i\gamma,\big) f(\eta_i = w_i - z_i\gamma)$ with $f(\cdot)$ the density of η_i, and $\Pr(L_i = 1 | x_i, z_{i,} y_i) = \Pr(\xi_{it} < -\xi_{it}^*(x_i, z_{i,}, y_i))$.[26]

To complete the parameterization, assume that ϵ and η are joint normal with variance-covariance matrix, $\Lambda = \begin{pmatrix} \sigma_\epsilon^2 & \cdot \\ \sigma_{\epsilon\eta} & \sigma_\eta^2 \end{pmatrix}$. The parameters of the model to be estimated include $\beta, \gamma, \lambda, \sigma_\epsilon^2, \sigma_\eta^2$, and $\sigma_{\epsilon\eta}$. As is well known, joint normality is sufficient to identify the wage parameters (γ and σ_η^2) as well as $\frac{(\sigma_\eta^2 - \sigma_{\epsilon\eta})}{\sigma_\epsilon^2}$ (Heckman 1979). With the exclusion restriction that there is a variable in x that is not in z, identification doesn't have to rely solely on the distributional assumption. The data on work choices identify $\beta/\sigma_\xi, \gamma/\sigma_\xi$ and λ/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 z and x. Having identified the parameters of the wage function (the γ's), the identification of σ_ξ (and thus also $\sigma_{\epsilon\eta}$) requires the existence of at least one variable that appears only in the wage equation, a variable in z and not in x. With that exclusion restriction, all of the elements of ξ_i^* are identified.

The identification argument is independent of the existence of the welfare program. That is, the model parameters can be identified from data either with or without the program in place. With parameter estimates in hand, the ex ante impact of the welfare program on employment, $\Pr\big(L_i = 0 | x_i, z_i, n_i, y_i, b(y_i, n_i)\big) - \Pr(L_i = 0 | x_i, z_i, n_i, y_i, b(y_i, n_i) = 0)$, can be obtained for various welfare benefit schedules $b(y_i, n_i)$.

To understand the contribution of the parametric model, note that the hypothetical program considered above excluded working women from eligibility. Suppose, more realistically, that the program allows working women to receive welfare benefits, but that women who work are subject to reduced benefits that depend on their earnings. Specifically, assume that the there is a benefit reduction (tax) rate that is proportional to earnings and that net benefits are given by $b(n_i, y_i) - \tau(n_i)w_i \geq 0$, where $\tau(n_i)$, the tax rate on earnings, depends on the number of

[26] The number of children only enters the model through the welfare schedule. Allowing for either a preference or cost of children, and assuming fertility is not a choice, does not change the conclusions from the analysis. As shown in Wolpin (2013), nonparametric ex ante evaluation in not feasible if fertility is a choice.

children. The budget constraint in this case is

$$(10)\quad \begin{aligned} C_i &= y_i + w_i(1 - L_i) \\ &\quad + \big(b(y_i, n_i) - \tau(n_i)w_i\big)L_i, \\ &= \big(y_i + b(y_i, n_i)\big) \\ &\quad + \Big(w\big(1 + \tau(n_i)\big) - b\Big)(1 - L_i) \\ &\quad - \tau(n_i)w_i, \\ &= \tilde{y}_i + \tilde{\tilde{w}}_i(1 - L_i) - \tau(n_i)w_i. \end{aligned}$$

Clearly, the form of the budget constraint no longer conforms to the case without the welfare program. Nonparametric estimation of the ex ante program effect using the previously described matching estimator is infeasible.

On the other hand, the parametric model parameters can be estimated in the absence of any data on the welfare program and the model can be used to assess the policy effects of the welfare program with the benefit reduction tax. The latent index governing labor supply decisions is given by

$$(11)\quad \begin{aligned} & v_i^*(x_i, w_i, \eta_i, \epsilon_i) \\ &= -\lambda y_i - (1 + \lambda)b(y_i, n_i)\,) \\ &\quad + z_i\gamma((1 + \lambda)\tau(n_i) + 1\,) - x_i\beta \\ &\quad + ((1 + \lambda)\tau(n_i) + 1\,)\eta_i - \epsilon_i \\ &= \xi_i^{\,*} + \xi_i. \end{aligned}$$

A policy maker can be provided with a menu of options that vary the benefit schedule and tax rate. Using the estimated model, it is possible to perform an ex ante evaluation of their effects on employment, take-up rates, and costs.

Most of the literature we review adopts parametric models, either static or dynamic, of individuals' decision-making processes. In the context of the previously described model, one dynamic extension would be to allow the wage offer function to depend on prior work experience. An additional way of extending the model might include additional choices, such as schooling, fertility and marriage.[27] Assuming discrete time, and that the woman maximizes discounted expected lifetime utility, and that future realizations of preferences and wage offers are unknown, the decision problem involves solving a discrete choice dynamic programming (DCDP) problem. There are now a number of survey articles that provide detailed discussions of available methods for estimating DCDP models (See Keane, Todd, and Wolpin 2011).[28]

3.2 *Incorporating an RCT*

Suppose a government is contemplating the introduction of a welfare program. To better understand the program's impact on female employment, the government decides to do an RCT. Given the cost of conducting an RCT, the government chooses only one benefit schedule, $b(n_i, y_i)$, and sets $\tau(n_i) = 0$. The sampling frame includes all unmarried women with at least one child, independent of their employment status. Women are randomized into two groups, one of which is offered the program, the treatment group, and one of which is not, the control group. In addition, the government collects data on the women's wage histories, unearned income, fertility, marital status and employment. The experimental impact estimates show a signifi-

[27] For example, see Keane and Wolpin (2010).

[28] Dynamic models require an explicit assumption about whether a policy change is anticipated. Although there are a few exceptions, the literature has generally assumed policy changes to have been a surprise.

cantly lower employment rate after one year for women in the treatment group.

After completing the RCT, the government makes the data available to researchers. Given that the treatment effect has already been calculated (including for subgroups based on observable characteristics collected in the survey, e.g., race, education, employment histories, etc.), some researchers decide the data offer nothing more to study. They advise the government to do additional RCTs to study the impact of varying the benefit schedule. Other researchers begin work on developing estimable models for the purpose of evaluating variations in the program's design.

The latter researchers have decisions to make with regard to model specification and estimation sample. Model selection is often done through a process by which a researcher tries to improve the model fit during a model-building phase, iteratively altering the model structure and reassessing within-sample model fit. This process is sometimes referred to as data mining, and it carries with it the dangers of overparameterizing the model to fit the data. Given this process, it can happen that models with different structures fit the data equally well. Conventional standard errors are also incorrect if they do not account for the iterative model selection process.

An alternative to using within-sample fit statistics in selecting the best model is to use a holdout sample and to look at an out-of-sample fit criterion.[29] To make decisions about whether to withhold some of the data in estimation and which data to withhold, the researcher should have a model in mind. To see why, consider the previous model of welfare participation decisions augmented to include a direct effect of welfare participation on utility, that is, a stigma effect associated with program take-up. Specifically, let the utility function be given by

$$(12) \quad U_i = C_i + \alpha_i L_i + \lambda C_i L_i - \varphi_i P_i,$$

where $P_i = 1$ indicates that the woman takes up the program (conditional on eligibility), $P_i = 0$ if she does not, and $\varphi_i = \overline{\varphi} + \omega_i$ is the woman's psychic disutility of participating in the welfare program (stigma). To make the point most clearly, assume that the program only applies to nonworking women. In that case, the budget constraint is

$$(13) \quad C_i = y_i + w_i(1 - L_i) + b(y_i, n_i)\,P_i.$$

The choice set for an eligible woman is now work, $L_i = 0$, not work and take up the program, $L_i = 1$ and $P_i = 1$, or not work and not take up the program, $L_i = 1$ and $P_i = 0$. The alternative-specific utilities are:

$$(14) \quad U_i(L_i = 0) = y_i + z_i \gamma + \eta_i,$$

$$(15) \quad U_i(L_i = 1, P_i = 1) = (1 + \lambda)\big(y_i + b(y_i, n_i)\big) - \overline{\varphi} - \omega_i + x_i \beta + \epsilon_i,$$

$$(16) \quad U_i(L_i = 1, P_i = 0) = (1 + \lambda) y_i + x_i \beta + \epsilon_i.$$

As can be seen, the stigma effect $\overline{\varphi}$ is identified from the proportion of women who are eligible for the welfare program but choose not to take it.[30] It is clear that the stigma

[29] From either a Bayesian or classical perspective, absent data mining there is no rigorous rationale for holding out data. In the Bayesian case, the marginal likelihood carries with it a penalty for models with more parameters (see Schorfheide and Wolpin 2012 for a discussion of these issues) and, in the classical case, one can adopt a degrees-of-freedom penalty function such as the Akaike or Bayesian information criterion.

[30] A woman will not take up welfare, $P_i = 0$, if $\omega_i \geq (1 + \lambda) b(y_i, n_i) - \overline{\varphi}$, and will take it up otherwise. Note that evidence for the existence of stigma based on

effect cannot be identified using only control group data, and that estimating the model using control group data alone will not generate accurate forecasts of program effects without good a priori evidence on $\overline{\varphi}$.[31]

The fact that the stigma parameter is identified from the take-up rate of the treatment group has implications for the choice of the holdout sample. If the model is estimated on the treatment group, the labor supply behavior of the control group, which is not subject to the program, can be simulated and compared to the data. Estimation based on the treatment group allows for counterfactual welfare policies to be simulated under the assumption that the stigma effect is not altered under these policy changes. As will be seen below, holding out the control group has sometimes been a validation strategy adopted in the literature.

If a researcher commits to holding out either the treatment or the control group, all data mining in terms of model development must be based only on the estimation subsample. If all the data are used for estimation, then the opportunity for out-of-sample validation is eschewed. As our review of this literature demonstrates, there does not seem to be a consensus yet, certainly on the choice of models but also on the best choice of estimation/holdout sample. However, much evidence has accumulated on the performance of different kinds of models and validation approaches.

4. *Model Validation*

As illustrated above, a major benefit of a structural modeling approach is that it allows for ex ante evaluation of policy interventions as well as consideration of alternative policy designs and eligibility criteria. However, models typically rely on extra-theoretic modeling and distributional assumptions, so model validation is an important concern.

4.1 *Approaches to Assessing Model Validity*

There are primarily three different approaches that researchers take to assess model validity. The first is to check robustness to alternative modeling assumptions, which was Leamer's (1983) suggestion. This requires estimating many different versions of the model and comparing the results obtained, which, especially in the type of estimation problems considered here, can be computationally intensive.

A second traditional way of considering model validity is to examine within-sample fit. Once the model parameters are estimated, including the parameters of the distributions of any unobservables, the estimated model can be used to simulate the choices and outcomes of individuals. To examine the within-sample model fit, one compares the actual choices and outcomes observed in the data to those simulated under the model. Formal within-sample fit tests can be conducted (for example, a Pearson chi-square test).[32] Such tests, however, are biased

"eligible" women not taking up the program relies on there not being significant measurement error in the data used to infer eligibility. Women classified as eligible may be observed not to take up the program because they are in fact not eligible, which could rationalize a model in which there is no stigma. It would be possible to estimate the classification error only using the treatment group data, in which case the control group could serve as a holdout sample.

[31] Nonparametric identification of the wage offer function requires an exclusion restriction, a variable that shifts preferences (x) that does not shift the wage (conditional on the z-variables). If the researcher does not have a plausible exclusion restriction and does not want to rely solely on distributional assumptions for identification, then the wage offer function could also be identified by making use of the randomized treatment variation. In that case, however, the researcher uses all the data in estimation and forgoes the opportunity of using a holdout sample for model validation. There is also an important caveat; if there are general equilibrium effects on wages due to employment effects of the program, then the treatment itself affects wages, that is, it is z-variable and cannot serve as an exclusion restriction.

[32] In the context of structural estimation, it is formally necessary to adjust degrees of freedom of the test for

toward not rejecting the model when the researcher engaged in data mining.

A third way of evaluating a model's validity is to use a holdout sample. Under this approach, the model is estimated on a subsample of the data and then used to predict the behavior of the holdout sample. In the case of an RCT, the use of a holdout sample as a validation tool has strong intitutive appeal. The RCT alters the structure of the decision problem faced by the agents in the treatment group and simultaneously ensures that distribution of observables and unobservables are the "same" across treatment and control groups. Depending on whether the conditions for identification are satisfied, it may be possible to recover the model parameters using only the control (or treatment) group. To be able to accurately forecast the reaction of agents to the treatment based on data from either the control or treatment samples alone is a nontrivial test of the model that possibly provides a basis for selecting among (or combining) models.

To our knowledge, Schorfheide and Wolpin's (2016) (hereafter SW) is the only paper to go beyond the intuitive argument and provide a formal justification for the use of a holdout sample. Their approach is to cast the problem of model selection as a principal–agent problem. A policy maker, the principal, would like to predict the effects of a treatment at varying treatment levels. The data are available to the policy maker from an RCT that has been conducted for a single treatment level. To assess the impact of alternative treatments, the policy maker engages two modelers, the agents, each of whom estimates their preferred structural model and provides measures of predictive fit.

Modelers are rewarded in terms of model fit. SW consider two data venues available to the policy maker. In the first, the no-holdout venue, the modelers have access to the full sample of observations and are evaluated based on the marginal likelihood function they report, which, in a Bayesian framework, is used to update model probabilities. Because the modelers have access to the full sample, there is an incentive to modify their model specifications and thus overstate the marginal likelihood values. SW refer to this behavior as data mining. More specifically, data mining takes the form of data-based modifications of the prior distributions used to obtain posteriors. In the second, the holdout venue, on the other hand, the modelers have access only to a subset of observations and are asked by the policy maker to predict features of the sample that is held out for model evaluation. Data mining creates a trade-off between providing the full sample, which would otherwise be optimal for prediction, and withholding data. SW provide a qualitative characterization of the behavior of the modelers under the two venues based on analytical derivations and use a numerical example to illustrate how the size and the composition (in terms of observations from the control and treatment groups) of the holdout sample affects the risk of the policy maker. Their numerical example shows that it is possible for the holdout venue to dominate the no-holdout venue because of the data mining that occurs if the modelers have access to the full sample. The lowest level of risk in their example is attained by holding back 50 percent of the sample (where the control and treatment sample are of equal size) and providing the modelers only with data either from the control or from the treatment group.

4.2 *Model Transparency*

Andrews, Gentzkow, and Shapiro (2017, 2020, henceforth ASG) propose that structural models be evaluated on the basis of a new "transparency" criterion that they define. They describe a scenario where a

estimated parameters. See Heckman (1984) and Andrews (1988).

reader (e.g., policy maker) has to make a decision based on statistics a researcher provides. The researcher generates a parameter estimate under an assumption a_0 and presents that estimate along with other data-derived statistics. The decision maker, however, is concerned about possible misspecification and considers a range of alternative possible assumptions $a \in A$. ASG define transparency as the relative reduction in the expected loss function from basing the decision on the researcher's supplied statistics relative to using the full dataset.

Bonhomme (2020), in a comment on ASG's (2020) paper, expresses skepticism about the usefulness of ASG's transparency criterion for counterfactual policy analysis. He emphasizes that while the transparency criterion can be helpful for understanding how the researchers' modeling assumptions influence model estimates, it is likely to be considerably less informative for understanding the reliability of model predictions that are outside the range of the data. When models are used for out-of-sample prediction, and particularly for counterfactual policy evaluation, as is often the goal of the structural approach, Bonhomme suggests validation based on holdout samples as a complementary method for achieving greater transparency.[33]

[33] ASG use the following counterfactual policy experiment (performed by both Attanasio, Meghir, and Santiago (2012) and Todd and Wolpin (2006)), seemingly a counterexample of how descriptive statistics can provide transparency for counterfactual policy evaluation. In the counterfactual, the school attendance subsidy is eliminated for the youngest children, who almost universally attend school, and the program cost savings are redistributed as a larger subsidy for older children. To the extent that older children's school attendance is responsive to the higher subsidy, overall school attendance increases. ASG posit that the difference in the treatment effects for younger versus older children estimated under the RCT is revealing of the new subsidy schedule impacts. However, the critical information needed to perform the counterfactual is how older children respond to increased subsidy amounts, for which there is no obvious set of descriptive statistics to make a judgment about the reliability of the model's prediction.

In sections five and six, we review papers from the new literature that combines structural modeling with data from RCTs or from quasi-experiments. Some of the studies use all the data in estimation and some exclude either the treatment or control group for use as a holdout sample. The estimated models are used for various purposes. Most studies use the model estimates to evaluate the effects of policies that deviate in some ways from the policy that was implemented, as described in the previous example. However, some papers are also concerned with spillover effects from treated individuals onto untreated individuals, or with general equilibrium effects arising from demand- and supply-side market responses, and they develop modeling frameworks to account for these possibilities.

5. *Applications*

5.1 *Conditional Cash Transfer Programs*

As previously described, one class of programs that has been studied using the structural approach is conditional cash transfer (CCT) programs, particularly in the area of education. We first describe two dynamic models that were developed and estimated to study the effects of the Progresa CCT program in rural Mexico on schooling, labor supply and fertility outcomes. Then we describe a simpler static model that was also used to study impacts of CCT programs in Mexico and Ecuador on school and child work choices. Third, we describe a model that was developed to study teacher attendance decisions in India and to analyze the effect of a teacher attendance subsidy and bonus program. These three studies exploit RCT data in different ways to estimate and validate structural models and then use the models to perform a range of counterfactual experiments. Lastly, we describe a study of the Progresa CCT program that uses quasi-experimental data from urban areas to study food demand.

5.1.1 *RCT Studies*

Effects of the Progresa program on schooling and fertility outcomes. In 1997, the Mexican government introduced a conditional cash transfer (CCT) program in rural areas that provided a subsidy to families for each child regularly attending school. The initial program, called Progresa, was afterwards extended to urban areas (and renamed Oportunidades and later Prospera). Similar programs have been adopted in numerous other countries (for example, in Bangladesh, Brazil, Colombia, Guatemala, Malawi, Nicaragua, and Pakistan).

To evaluate the initial program, the Mexican government conducted a randomized social experiment in which 506 rural villages were randomly assigned to either participate in the program or serve as controls. Randomization, under ideal conditions, allows mean program impacts to be assessed through simple comparisons of outcomes for treatments and controls. The program was effective in increasing school attendance; treatment effects, measured as the difference in average attendance rates of children in the treatment and control villages one year after the program, ranged from 5 to 15 percentage points depending on age and sex (Behrman, Sengupta, and Todd 2005; Schultz 2004).

An important limitation of large scale social experiments such as Progresa is that it is often prohibitively costly to vary the experimental treatments in a way that permits evaluation of a variety of policies of interest. In the Progresa experiment, all eligible treatment group households faced the same subsidy schedule, so it is not possible to evaluate the effects of alternative subsidy schemes through simple treatment-control comparisons.[34] In addition, because the experiment lasted only two years, one cannot directly assess long-term program impacts.

Todd and Wolpin (2006, hereafter TW) and Attanasio, Meghir, and Santiago (2012, hereafter AMS) analyze the impact of the Progresa program on school attendance via the estimation of a DCDP model of decision-making about children's schooling. They use their model estimates to compare the effects of the existing program to the effects of various alternative program designs. Both papers adopt the DCDP approach, use data derived from the same source, and perform similar counterfactual exercises; however, the models used differ nontrivially in their structure.

We first provide a general description of the Progresa data and then describe the two models, their different approaches to using the data, and their empirical findings. A baseline survey was conducted in October 1997 of all households in both the treatment and control villages prior to the program's implementation. The experiment began in the 1998/99 school year and continued for two years.[35] The program (which included a child health component as well) provided benefits that, on average, amounted to about 25 percent of family income. The school attendance subsidy component amounted to about 75 percent of total payments. The subsidy began at grade three and increased with each additional completed year of schooling to offset the increased opportunity cost of attending school as children become older. The subsidy level was the same for girls and boys up to grade six, but was larger for girls in grades seven to nine.

In the TW model, a married couple decides in each year whether each of their children between the ages of 6 and 15 will attend school, remain at home or, for those

[34] Under Mexican law, it was illegal to offer different subsidy schedules to different eligible families.

[35] Within the treatment villages, only households that satisfied an eligibility criterion based on a "marginality" index were provided with the subsidy.

age 12 to 15, work in the labor market (the choices are mutually exclusive). They also decide whether the wife will become pregnant (while fecund). The couple receives utility in each period from their stock of children, their children's current years of schooling, their school attendance, and from any children at home. There is also a utility cost to attending school (grades seven–nine) that depends on the distance from the village to a school. Households differ in their preferences for the choice variables according to their discrete unobserved "type," and households have time-varying preference shocks (normally distributed). The household's income includes the parents' income and the wage income of the children who work.[36] Model parameters are estimated by simulated maximum likelihood.

The AMS model also includes the binary choice of school or work (excluding the "at home" option), but, unlike the TW model, assumes that each child's utility is maximized independently of that of the parents or of other children. 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. The child receives a wage offer in each period that is village/education/age-specific. If the child rejects the wage offer and attends school, the child receives a utility payoff (positive or negative) that depends on observable preference shifters (parental background, the child's age, and the state of residence), the number of years of past attendance, on observable variables that affect the cost of attending primary or secondary school (distance to a secondary school), on a child's unobserved discrete preference "type," and on a time-varying preference shock (distributed type I extreme value).

The AMS model is consistent with a direct effect of the program on school attendance utility, either a "feel good" effect from participating in the program or a "stigma" effect. Unlike in the welfare example where individuals may decide to work and not take up welfare, all families in the Progresa experiment received the subsidy if their child attended school. Thus, the possibility that there may be an intrinsic value of program participation per se would require that both the treatment and control households are used in estimation. AMS use both groups in estimation. As Wolpin (2013) points out, because AMS do not fully specify the constrained optimization problem, it turns out that their model is observationally equivalent to one in which there is no direct program effect on utility. Thus, estimation of the partial equilibrium decision model did not strictly require both the treatment and control groups. In contrast, TW, assuming that there is no intrinsic value of participation, hold out treated households in estimating the model, using these households instead for purposes of out-of-sample model validation.

TW compare the predicted effects of the Progresa program on completed schooling, as implemented, with that of alternative programs. Model simulations of households from the time of marriage until the last born child reaches age 16 show 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 model predicts an increase in completed schooling of about one-half of a year for both boys and girls, or 26.0 percent of the maximal potential increase for girls and 28.9 percent for boys.[37]

[36] A child's wage (offer) depends on the child's age and sex, the distance to the nearest city, household type, and unobserved shocks.

[37] Interestingly, this estimate corresponds closely with that obtained by Behrman, Sengupta, and Todd (2005) and Schultz (2004) using nonstructural approaches.

As noted, the Progresa subsidy schedule rewards school attendance starting at grade three. However, attendance in grades three–five is almost universal, making the subsidy at early grade levels essentially an income transfer. TW calculated that the per family cost of the program could be held roughly constant if the subsidy in grades three–five were eliminated and the subsidy in grades six–nine were increased by about 45 percent. Under the modified program design, the proportion of girls completing ninth grade increases by 3.4 percentage points and proportion of boys by 3.8 percent, although there was a small decline in the proportion of children who complete at least sixth grade. TW also use the model to evaluate alternative hypothetical programs, such as a bonus for completing ninth grade, a school building program that decreases the distances that students need to travel to attend school, and an unconditional family income transfer program.

AMS perform two counterfactuals. As in TW, they simulate the impact of eliminating the subsidy to primary school and redistributing the savings to increase the subsidies at later grades, and they simulate the impact of building schools. Like TW, they find the effect of the first counterfactual to be large, although the metric used by AMS is not directly comparable to that of TW. They find that the budget-neutral effect of eliminating the subsidy at younger ages increases age 15–16 school attendance rates by as much as 100 percent. Also consistent with TW, AMS find a large effect of building schools on older children's school attendance. The TW and AMS findings are, perhaps, surprisingly similar given the quite significant differences between the model structures and estimation samples.[38]

Effects of CCTs on schooling and work in Mexico and Ecuador. A study by Leite, Narayan, and Skoufias (2015) uses microsimulation methods to perform an ex ante evaluation of CCTprogram impacts on school enrollment and child working. The model they specify is based on a model originally developed in Bourguignon, Ferreira, and Leite (2003) in studying effects of the Bolsa Escola CCT in Brazil. The model is a static discrete choice random utility model where the options for each child are to not attend school, to combine schooling and working, or to only attend school. The model assumes that decisions to send a child to school are independent of parents' working decisions, that decisions about multiple siblings are made independently, and that family composition is exogenous. Utility depends on family income, inclusive of child wages, and any program transfers associated with the alternative school/work choice combinations. The model incorporates a "means test" to approximate program eligibility.

Estimation of the model does not require panel data and is much less demanding in terms of computational complexity than the TW and AMS models described above. Nevertheless, when Leite, Narayan, and Skoufias (2015) compare the ex ante predictions from the model to experimental benchmark estimates from the Mexican Progresa experiment and the Bono de Desarollo program in Ecuador, they find that the model produces reliable forecasts.[39]

[38] As reported in Wolpin (2013), the predicted effect of doubling the subsidy, a large out-of-sample change for both the AMS and TW model, was also quite similar in the two studies.

[39] Under the Bono de Desarollo program, beneficiary households receive grants of \$15 per month under the conditions that children of ages 6–16 years are regularly enrolled in school with an attendance rate of at least 80 percent per month and children of ages 0–5 years make scheduled visits to health centers. Coverage reached one million households (5 million people). Schady et al. (2008) present experiment impact estimates showing positive effects on school enrollment and negative effects on child work.

Effects of teacher attendance subsidies in India. Duflo, Hanna, and Ryan (2012) analyze the impact of financial incentives and teacher attendance monitoring on teacher absenteeism in rural India. In September 2003, an NGO implemented an RCT that randomly assigned 60 of 120 schools to a treatment group in which teacher monthly salaries were determined by a nonlinear function of the days per month that they attended school. Treatment group attendance was monitored by requiring a photograph be taken of the teacher and students at the beginning and end of each school day using a camera with tamper-proof time and date functions. The salary structure consisted of a flat payment for attending 20 days in the month, a 5 percent bonus payment for each day above 20 (about 3 days per-month on average), and a 5 percent penalty for each day below 20 (up to 10 days missed). Teachers in the control group schools faced the same flat payment, with neither a bonus for additional days above 20 nor a penalty for days fewer than 20. Attendance was monitored through random monthly checks, and control group teachers were reminded that they could be fired for excessive absences.

Duflo, Hanna, and Ryan (2012) specify a finite horizon DCDP model of the teacher's daily attendance decision. They estimate different specifications, including observed and unobserved preference heterogeneity, iid preference shocks, and serially correlated preference shocks. As seen in table 1, they estimate the model on the treatment group and use the control group to select and validate the model specification. The treatment consisted essentially of two bundled treatments, the financial incentive and the use of the camera for monitoring absences. Estimating the model on the treatment group and predicting the behavior of the control group might have led to an overstatement of attendance of teachers in the control group if there was an additional effect of the camera monitoring technology. However, the authors find that the model forecasts are accurate and conclude that the camera monitoring had little effect above that of the incentive payments.[40]

As noted, the RCT included only one form of financial incentives. However, it is possible, given model estimates, to calculate the optimal incentive scheme, that is, the financial incentive structure that produces the same absentee rate at least cost. When the authors use the model to find the optimal incentive structure, they find that the optimal structure saves 22 percent of the average cost associated with the incentive structure implemented in the experiment.

The papers in this section illustrate different ways that RCTs have been combined with structural modeling, representing different judgments by researchers about the value of a holdout sample in model validation relative to the using the exogenous variation induced by the RCT in estimation. The TW and Duflo, Hanna, and Ryan (2012) studies both use a holdout sample. TW and Duflo, Hanna, and Ryan (2012) could, in principle, could have held out either the control or treatment group. In contrast, AMS and Leite, Narayan, and Skoufias (2015) used both treatment and control group data in estimation.[41]

[40] The validation exercise identified two specifications that gave similar out-of-sample performance.

[41] There are some very recent papers combining RCTs and structural modeling of college attendance decisions. Tincani, Kosse, and Miglino (2022) combine RCT data and structural modeling to analyze the impacts of a preferential college admissions program in Chile (called PACE) that guarantees university admission to students with GPAs in the top 15 percent of the class. The study uses both the control and treated groups in estimating the model, with the goal of the modeling being to better understand the mechanisms underlying the observed treatment effects. The RCT showed negative impacts on high school study effort, which the model estimates show is partly attributable to biased beliefs that students have about their ranking in the overall distribution. Belzil, Pernaudet, and Poinas (2021) use both survey data and an RCT to elicit high school students' valuations of college loans and financial aid. The survey data is used to estimate a structural model

TABLE 1
STUDIES OF CCT PROGRAMS IN EDUCATION

Study	Group used for estimation	Out-of-sample model validation?	Evaluate counterfactual programs?
Todd and Wolpin (2006) Model of school going, child labor, and fertility used to evaluate effect of CCTs in Mexico	Control	Yes, using treated sample	Yes, different subsidy designs compulsory schooling laws, child labor law enforcement
Attanasio, Meghir, and Santiago (2012) Model of school going and child labor in Mexico used to evaluate effect of CCTs in Mexico	Control & treatment	No	Yes, different subsidy designs
Leite, Narayan, and Skoufias (2015) Model of school-going and child work used to evaluate effect of CCTs in Mexico and Ecuador	Control	Yes, using treated sample	Yes, different subsidy designs
Duflo, Hanna, and Ryan (2012) Model of teacher attendance in India used to evaluate effect of teacher subsidies	Treatment	Yes, using control sample	Yes, different incentive schemes
Angelucci and Attansio (2013) Engel curve model of food demand	Control and pre-reform treatment	Yes, compare to quasi-experimental diff-diff and matching estimates	No

5.1.2 *Quasi-experimental Studies*

Effects of the Progresa program on food demand. Angelucci and Attanasio (2013) analyze the effect of the Progresa cash transfer program on food demand in urban areas of Mexico. Their data come from a quasi-experiment that made the program available to households in some localities but not in others. The authors use two different evaluation estimators—a propensity score matching estimator and a difference-in-difference estimator—applied to longitudinal data from households in treated localities and matched control localities. The matching estimates show that the program led to an increase in the food expenditure share and an increase in high-protein food consumption.

One of the goals of the paper is to assess whether a standard Engel curve model could be used to do an ex ante prediction of the program impacts. The Engel curve relates food expenditure shares and high-protein food expenditure shares to total expenditure. The authors estimate Engel curve demand

of college-going decisions, incorporating students' beliefs about whether they will attend college, from which valuations of financial aid and student loans are inferred. In the RCT, the same students are offered choices between cash options and future financial aid. The authors conclude that there is an incoherency between the student evaluations elicited in the survey and those obtained from the field experiment.

models using data on control households collected at times before and after the program and on treatment households collected prior to the program. The parameter estimates indicate that food is a necessity and high-protein foods are a luxury.

The impact estimates based on the matching and difference-in-difference estimators showed the treatment group increased their expenditure share on food, which is inconsistent with the decline predicted by the estimated Engel curve. For high-protein food, the quasi-experimental evidence and the estimated Engel curve both predict an increase, although the treatment effect estimate is larger in magnitude than that predicted by the Engel curve. When the Engel curve is estimated separately on the treatment group before and after the program, the parameter estimates change substantially, which the authors interpret as additional evidence of model misspecification.

The authors hypothesize that the Engel curve, which represents consumption demand of a unitary household, does not account for the fact that the Progresa cash transfers were given to women and that decision-making within the household may rather be the result of a bargaining process. When they reestimate the Engel curve only on the subset of single female-headed households, they find that the parameter estimates using treatment group data from before and after the program are stable, supporting their conjecture about the source of model misspecification.

5.2 *Welfare Programs*

5.2.1 *RCT Studies*

Effects of cash transfers in Indonesia. Alatas et al. (2016) analyze the effects of a welfare (cash transfer) program in Indonesia called PHK, specifically, how the mechanism that is used to enroll people affects program take-up rates and impacts. PHK enrolled 2.4 million households, each receiving $130 per year for six years, with eligibility determined on the basis of an asset test. The authors carried out an RCT that varied the enrollment process across villages. In 200 treated villages, they introduced a "self-targeting" scheme, whereby households had to travel to apply for the program at a registration site and to take an asset test to determine eligibility. The RCT also randomly varied the application costs across treated households by varying the distance needed to travel to the registration site. In 200 control villages, they followed the usual government "automatic screening" procedure with program administrators visiting potential beneficiaries at their homes to determine eligibility.

The RCT revealed that the different enrollment schemes result in very different patterns of program participation. Per capita household consumption is lower for participating households in the treated villages than in the control villages. In fact, the very poorest households, as measured by per capita consumption, were twice as likely to receive benefits under the self-targeting scheme. However, only about 60 percent of eligible households apply under self-targeting, so the program coverage rate is lower.

To better understand the mechanisms generating the different enrollment patterns, the authors develop and estimate a discrete choice model of the household's program application decision under uncertainty about whether they will they will pass the asset test. In the model, households weigh the expected benefits of applying against the costs, inclusive of any distance travel costs. The model incorporates two types of households—sophisticated and unsophisticated—with sophisticated households being better informed about the income components that comprise the asset-based eligibility test. As seen in table 2, the discrete choice model is estimated using the treatment group data,

TABLE 2
STUDIES OF WELFARE PROGRAMS

Study	Group used for estimation	Out-of-sample model validation?	Evaluate counterfactual programs?
Atalas et al. (2016) Model of decision to apply to a subsidy program in Indonesia that is used to evaluate different targeting mechanisms	Treatment	No	Yes, change program application costs (time, distance), change expected prob. of receiving benefits, and change fraction well-informed about eligbility rules
Card and Hyslop (2005) Logistic panel model of welfare participation in Canada used to decompose effects of income supplement	Control	No	No
Lise, Seitz, and Smith (2015) Job search and matching model in Canada used to evaluate effect of income supplement welfare program	Control	Yes using treated sample	No
Choi (2018) Model of labor supply and welfare participation in US used to evaluate effect of changing benefit reduction rates	Control	Yes using two treated samples	No
Keane and Wolpin (2007, 2010) Model of labor supply, fertility, welfare participation in US used to evaluate effect of changes in welfare rules	Control	Yes using held-out state	Yes, changes in welfare benefits
Hansen and Liu (2015) Model of labor supply and welfare participation in Canada used to evaluate effect of changes in welfare rules	Pre-reform treatment and control	Yes using post-reform treated	Yes, changes in welfare benefits and income tax schedule

and model fit is assessed with within-sample fit tests.[42]

Simulations from the estimated model show that a key factor driving the selection of poorer households into the program under self-targeting is that rich households forecast that they have a small likelihood of receiving benefits and therefore do not apply when there is an application cost. The estimates show that a small distance cost is effective in targeting the program to the poorest households and that further increasing the distance cost has no additional targeting benefit. The authors also use the estimated model to examine how application decisions

[42] Only treated households make the decision about whether to apply to the program.

change when the fraction of sophisticated households increases and when households change their expectation of receiving benefits. Lastly, they compare how the two types of enrollment schemes influence the poverty gap. They find that it is possible to achieve a 29 to 41 percent greater reduction in the poverty gap under self-targeting than under automatic screening with an identical budget.

Evaluating effects of an earnings supplement in Canada. Card and Hyslop (2005) and Lise, Seitz, and Smith (2015) use data from an RCT, the Canadian Self-Sufficiency Project (SSP), to analyze the effect of a wage subsidy given to long-term welfare recipients upon employment. The SSP provided an earnings supplement (a 50 percent negative income tax) for up to three years for individuals receiving Income Assistance (IA), the Canadian welfare program, if they obtained full-time employment within a 12-month time period.[43] As noted in Card and Hyslop (2005), the design of the program created different incentives. One is an incentive to gain employment quickly to establish eligibility for future subsidies, which they call the establishment incentive. The other is the entitlement incentive created by the negative income tax, which encouraged individuals to work rather than participate in IA. Card and Hyslop (2005) use a theoretical search framework to analyze how the program would be expected to affect reservation wages and entry and exit rates. Their empirical approach is to estimate a panel data model of the welfare entry and exit behavior without and with the SSP, rather than solving and explicitly estimating the parameters of the search model. They find that a dynamic logistic model with second-order state dependence provides a good within-sample fit in the control group data. They then augment the model to include treatment effects that represent the establishment and entitlement incentive impacts and a model of the SSP program eligibility process. The main goal of the empirical analysis is to decompose the treatment effects into the establishment and entitlement incentive components.

Lise, Seitz, and Smith (2015) use the same RCT data to calibrate a job search model, in the style of Pissarides (2000), only using data from the control group. The analysis is done separately for the provinces of New Brunswick and British Columbia, because the labor markets and unemployment benefits programs differ across provinces. The key model parameters are the discount factor, search friction parameters, and exogenous job separation rates. Second, they use the model to simulate the behavior of the treatment group and they compare the predictions with RCT estimates.[44] In particular, they examine outcomes related to job search intensity, job destruction, and earnings. Lastly, the authors recalibrate their model combining the control and treatment group data and examine resulting changes in the parameters and model fit.

In British Columbia, Lise, Seitz, and Smith (2015) find that the SSP impacts on the IA-to-work transition rates predicted by the model match very well the transition rates observed under the experiment. However, the predictions are less accurate for New Brunswick, for which the model predicts a higher transition rate than observed in the data. The study finds that the search effort

[43]The data contain information on 5,685 recipients: 2,827 control group members and 2,858 treatment group members. Lise, Seitz, and Smith (2015) focus on 3,346 single women who were regularly included in follow-up surveys.

[44]They simulate the behavior in partial equilibrium, because the experiment only affected a small subset of the economy and is therefore not expected to have equilibrium impacts.

cost must be higher in New Brunswick than in British Columbia to match the data. With regard to job destruction, the authors find support for the assumption of a constant job destruction rate, because there is no statistically significant difference in the employment survival rates for the treatment and control groups and also no change observed when the treatment group stops receiving supplemental payments. In terms of earnings, the hourly earnings rate did not differ for the treatment/control groups but the treatment group worked longer hours. Lastly, including the treatment group in calibrating the model changes the parameter estimates for New Brunswick but not for British Columbia.

Card and Hyslop (2005) and Lise, Seitz, and Smith (2015) represent two different empirical approaches to analyzing the same program with the same data. The Card and Hyslop (2005) study informs about the program impacts for the program as implemented, and provides insights into how the two different program incentives contribute to the observed impacts.[45] The structural framework adopted in the Lise, Seitz, and Smith (2015) study makes stronger modeling assumptions on the process governing dynamic job search and welfare program participation behaviors, but the model can be used to vary program eligibility rules as well as income subsidy levels.

Evaluating effects of welfare policy changes in Minnesota and Vermont. Choi (2018) uses data from two state welfare reform experiments conducted by Manpower Demonstration Research Corporation (MDRC) during the mid 1990s—the Minnesota Family Investment Project (MFIP) and the Vermont Welfare Restructuring Project (WRP)—to assess a structural model's performance in forecasting the effects of welfare rule changes. The paper develops and estimates static discrete choice models of labor supply and welfare participation that incorporate heterogeneity in preferences, fixed costs of working, and disutility associated with welfare take-up. The welfare policy impacts estimated under the RCT are used as a benchmark for the structural model predictions.

The model is a static labor supply/welfare participation model in which individuals face a finite and discrete set of choices.[46] The utility function is quadratic in hours and consumption and includes an interaction term (to allow consumption and leisure to be complements or substitutes). Consumption depends on earned income, taxes, the Earned Income Tax Credit (EITC), and welfare benefits.

In the two state experiments, individuals assigned to the control group received the standard Aid to Families with Dependent Children (AFDC) program, which has a 100 percent welfare benefit reduction rate for every dollar earned. Individuals assigned to the treatment groups faced lower benefit reduction rates—62 percent in Minnesota and 75 percent in Vermont. The lower benefit reduction rate generates an income effect and a wage effect and will increase work if the wage effect dominates.

The MFIP and WRP samples include 14,170 and 7,691 individuals in the three program groups. Baseline survey data were collected prior to random assignment and in two follow-up surveys, 36 months and 42 months after random assignment. The

[45] Card and Hyslop's (2005) logit specification can be interpreted as an approximation to the solution of the behavioral model they present, an approach that might be called quasi-structural. However, because "deep" parameters are not recovered, it is not possible to analyze alternative program designs.

[46] The discrete choice assumption avoids the analytical difficulties created by nonlinear budget constraints with convex and nonconvex kinks. Similar models have been estimated by Fraker and Moffitt (1988), van Soest (1995), Hoynes (1996), Keane and Moffitt (1998), Gong and van Soest (2002), Creedy and Kalb (2005), Brewer et al. (2006), and Blundell and Shephard (2012).

model is estimated using data from the control group in Minnesota. Model parameters are identified from cross-section variation across individuals in hours of work and welfare participation. The stigma effect is identified because some eligible controls choose not to participate in AFDC.

After estimating six different model specifications using only the control group sample in Minnesota, Choi (2018) uses the estimated model to predict welfare policy impacts both in Minnesota (within-state) and in Vermont (cross-state). The RCT ensures that the distribution of unobservables for the control and treatment groups in Minnesota are comparable. However, performing the cross-state prediction requires an additional assumption that any unobservable factors governing labor supply and welfare participation decisions are similar in Minnesota and Vermont.

Choi (2018) finds that some of the model specifications provide a very good within-sample fit to labor supply and welfare participation patterns, particularly the specifications that incorporate a fixed cost of working. However, the model's out-of-sample predictions of the policy treatment effects are not good, either within state or cross-state. Specifically, the RCT estimates in Minnesota indicate that the decrease in the welfare benefit reduction rate induced a substantial decrease in hours of work, while the estimated models predict either a small decrease or an increase. In Vermont the RCT showed no change in welfare participation patterns, whereas the model predicts increases. The study concludes that a good within-sample fit is not necessarily indicative of good out-of-sample predictions. The results suggest that local labor market effects are potentially important in explaining heterogeneous program effects across regions and could not be adequately controlled in the cross-state forecasting analysis.

5.2.2 *Quasi-experimental Studies*

Evaluating effects of a welfare policy change in Canada. Hansen and Liu (2015) estimate a model of labor supply and welfare participation to perform an ex ante evaluation of a 1989 Canadian welfare reform. Prior to the reform, welfare benefits were much less generous for people younger than 30 years of age than for similar people 30 or older. The reform eliminated age discrimination in benefit levels and increased the average monthly benefit for younger individuals from $185 to $507. The authors estimate a static discrete choice model where individuals choose among seven different hours of work options and whether to participate in welfare. The model also includes a stigma effect of welfare participation. It accounts for the detailed budget sets for each welfare work combination as well as the income tax structure. Model parameters are estimated by maximum likelihood using a sample of single men from Quebec collected prior to the reform (from the 1986 Canadian Census).

The authors perform an out-of-sample fit test of the model by comparing the model's predictions of the reform impacts to those obtained using a regression discontinuity design (RDD) estimator applied to post-reform data, exploiting the age discontinuity. They find that the estimated model predicts the employment reduction and the increase in welfare participation associated with the reform. The largest policy effects occur for lower-income individuals for whom there is a 4.5 percent decrease in employment and a 4.9 percent increase in welfare participation.

The authors also use the model to study how employment, welfare use, and hours of work would change as social assistance benefits are further increased. They find the responses to be highly nonlinear with respect to benefit increases. In addition, they use the model to explore how labor

TABLE 3
Studies of Early Childhood Programs

Study	Group used for estimation	Out-of-sample model validation?	Evaluate counterfactual programs?
Attanasio et al. (2020) Model of early child skill formation used to evaluate effect of home visitation program in Colombia	Control & treatment	No	Yes, analyze effects of different program components (home visits and program-induced changes in parental investments)
Rodriguez (2018) Model of labor supply and childcare choices used to evaluate effect of income and childcare subsidies in the US	Control & treated	Yes, to experimental moments not used in estimation	Yes, changes to subsidy design and conditionality requrements
Chan and Liu (2018) Model of female labor supply, childcare choices, fertility used to evaluate effect of home care subsidies in Norway	Control & treatment	No	Yes, tax policies maternal leave

supply and welfare participation changes in response to changes in the income tax system.

Evaluating effects of welfare policy among states in the United States. There is a large literature on structurally estimating models to assess the impact of welfare programs in the United States on economic and social outcomes. In the papers discussed previously, the holdout sample corresponded either to one of the RCT groups or to the treated group observed at a point in time prior to the program's implementation. In both cases, the treatment and control groups are thought of as comparable in terms of the sample distributions of unobservables.[47] Keane and Wolpin (2007) instead explicitly choose a nonrandom sample as the holdout sample, specifically a subsample with a considerably different level of treatment. In their case, the treatment level corresponds to the welfare program benefit generosity (AFDC), which varies across states. The holdout sample is a state (Texas) that, relative to the set of states in the "treatment" sample, provides considerably less generous benefits. The notion is that forecasting well the effect of a program far outside the range of the estimation sample program parameters should be a more demanding out-of-sample validation criterion. The authors conclude that their DCDP behavioral model produced plausible forecasts, more plausible than a purely statistical model. In their follow-up paper, Keane and Wolpin (2010) provide an analysis of the impact of the AFDC program and counterfactual policies on program take-up, labor supply, wages, fertility and marriage.

[47] Although in the case of Choi (2018), the holdout sample included both the control and treatment group in Vermont plus the treatment group in Minnesota.

5.3 *Early Childhood Programs*

5.3.1 *RCT studies*

A home visitation/parenting program in Colombia. Attanasio et al. (2020) study the effects of a randomized early childhood intervention in Columbia that was offered to households participating in the Columbian CCT program *Familias en Accion*. The intervention was targeted at children ages 12–24 months and consisted of weekly home visits (one hour per week) aimed at improving parenting skills and providing micronutrient supplementation.[48] The data were gathered by a household survey, by tests administered to the children, and by interviewer observations. In total, 1,429 children were randomized into four groups: (i) one group that received only the psychosocial stimulation program, (ii) one group that received only the micronutrient intervention, (iii) one group that received both (i) and (ii), and (iv) a control group. Attanasio et al. (2018) report the RCT impact estimates that showed significant effects of the psychosocial intervention on child outcomes but no effects of micronutrient supplementation. Therefore, in Attanasio et al. (2020), groups (i) and (iii) and groups (ii) and (iv) are combined.

The primary goal of the Attanasio et al. (2020) study is to elucidate the mechanisms underlying the observed treatment impacts. To this end, the authors develop a model of the cognitive and socio-emotional skill production technology along with parental investment decision rules. The inputs in the production function model are baseline child skills, maternal skills, and material and quality time investments in the child. The production function also includes the presence of other siblings in the family who might reduce attention available for the focal child. The model incorporates a latent factor structure to combine multiple outcome and input measures and also allows for measurement error.[49] Some of the estimated specifications allow for material and time investments to be endogenous, using as instruments prices of toys and food and maternal exposure to violence.

The empirical analysis has two primary aims. The first is to understand the nature of the production function in this high-poverty context. The second is to ascertain whether the positive treatment effects occurred because of changes in the production function, changes in parental investment decisions, or changes in the mother's characteristics (e.g., rates of depression or socio-emotional skills).[50] The paper also decomposes production function changes into changes in total factor productivity, changes in other parameters, and a direct effect of the treatment, possibly operating through the one-hour home visits.

With regard to the skill production technology, the study finds that the current stock of cognitive (socio-emotional) skills strongly affects the development of future cognitive (socio-emotional) skills. This is called self-productivity of skills, using the terminology of Cunha, Heckman, and Schennach (2010). Second, the estimates show that the current stock of cognitive skills fosters the development of future socio-emotional skills, but not the reverse.

The treatment intervention increased children's cognitive development by 0.115

[48] This type of intervention was shown to be effective in the Jamaica Study (Grantham-McGregor et al. 1991) and in the Perry Preschool Program (Heckman et al. 2010). A difference in the Colombian program, however, was that the home visits were conducted by local women who received training but did not otherwise have expertise in child development.

[49] The approach is similar to that of Cunha, Heckman, and Schennach (2010).

[50] Because the treatment is allowed to affect model parameters, it is not possible to estimate the model using only the control group data and ex ante evaluation is not possible.

TABLE 4
Studies of Relocation/Migration Subsidy Programs

Study	Group used for estimation	Out-of-sample model validation?	Evaluate counterfactual programs?
Lagakos, Mobarak, and Waugh (2018) Model of urban-–rural migration in Bangladesh	Control & treatment	No	Yes, different amenities upon migration
Galiani, Murphy, and Pantano (2015) Discrete choice model of resid. location in Boston used to evaluate effect of conditional rent subsidy vouchers	Control and one treatment arm	Yes, using one treatment arm	Yes, alternative poverty threshold conditionality requirements

log points and socio-emotional development by 0.087 log points. The authors' preferred production function estimates imply that the parental investment increases (both in material and time) induced by the program account for around 91 percent of the intervention impact on cognition and at least 66 percent of its impact on socio-emotional skills. The parental investment increases were greater for children with higher initial baseline skills and for more highly skilled mothers. There is no evidence of a direct effect of the program and also no evidence that the program led to significant changes in the mothers' characteristics. The study concludes that the involvement of the parents and induced increases in parental investments were the key to the program's success.

An income and childcare subsidy program in Wisconsin. Welfare programs with work requirements often necessitate that parents make greater use of external childcare, raising concerns about how children are affected by such programs. Some of the best evidence on this issue comes from an RCT implemented by MDRC used to evaluate the New Hope program in Milwaukee, Wisconsin, and then also, in Rodriguez (2023), to study variations in the original program design. The RCT sample consisted of 1,357 individuals; 678 were randomly assigned to a treatment group and 679 to a control group. Data were collected from the families at baseline and up to eight years after. The treatment group received an income subsidy similar to the EITC and a childcare subsidy with a requirement to engage in full-time work. To be eligible, individuals had to be at least 18 years old and have a household income equal to or less than 150 percent of the federal poverty line. They received the subsidies for three years. The RCT showed significant positive program impacts on labor supply, family income, and childcare use. Interestingly, the RCT also revealed significant positive impacts on children's cognitive achievement. The treatment consisted of a bundle of conditional and unconditional subsidies and it is not possible to know from the RCT alone which of the components were most important in generating the positive impacts.

Rodriguez (2023) analyzes data from the New Hope RCT with the following goals: to understand the mechanisms that underlie the observed treatment impacts, to disentangle which of the program components was most important in generating the observed impacts, and to analyze impacts of modifying the program's design. The paper estimates a dynamic discrete choice model of the household labor supply and child human capital formation. In the model, a unitary household with a single child chooses hours of work and childcare types (informal home care or formal, center-based childcare). Household choices and the current stock of child human capital are inputs in the child human capital production function. The specification of the household's budget set accounts for different means-tested programs available to the household including AFDC, EITC, and New Hope. The model is estimated using a method of moments approach and only using nonexperimental moments. The model's predictions are compared to the experimental impact estimates.

The paper finds that New Hope's effects on child human capital are entirely explained by the childcare subsidy component, which led parents to take their children to center-based childcare. Model simulations show that giving an average family an amount of money equal to the cost of childcare increases child human capital by 0.8 percent of a standard deviation, but giving the same amount for restricted use in purchasing childcare services increases child human capital by 52 percent of a standard deviation. The greater productivity of external childcare in fostering human capital development accounts for the treatment effects that were observed on cognitive achievement. Rodriguez (2023) also uses the estimated model to evaluate the effects of varying the program design to not include the full-time work requirement, which he finds would lead to an even greater increase in children's human capital (by 0.04 standard deviations).

5.3.2 *Quasi-experimental Studies*

Childcare subsidy in Norway. Chan and Liu (2018), using data from a large-scale welfare reform in Norway, study effects of alternative childcare policies on women's life-cycle decisions and on long-term child cognitive outcomes. They develop and estimate a DCDP model of women's decisions with regard to labor supply, childcare, and fertility. The model allows children's cognitive development to be affected by childcare arrangements. The model is estimated using Norwegian administrative data that includes child test score data measured beyond age 10. The cognitive outcomes include scores on reading, math and English tests.

In estimation, the authors exploit a large-scale childcare reform called "cash for care," which provided cash to families with young children who did not use formal childcare options. They argue that this reform provides exogenous variation in the relative price of different childcare options that is useful to identify model parameters. The empirical results show that "cash for care" reform had a significant impact in reducing the employment rates of lower education mothers. The authors find that the use of nonmaternal early childcare leads to lower reading scores than formal care on average. The estimated DCDP model is also used to evaluate the effects of counterfactual policies, such as tax policies and maternal leave policies.

5.4 *Relocation/Migration subsidies*

5.4.1 *RCT Studies*

Housing subsidy in Boston. Galiani, Murphy, and Pantano (2015) study the effects of a housing rent subsidy on residential neighborhood choices. They use data from the Moving to Opportunity (MTO) housing subsidy experiment to estimate a model of household neighborhood choice and to analyze the effects of changing the program subsidy design. In the MTO experiment, low-income households in six cities (Baltimore, Boston, Chicago, Los Angeles, and New York City) were placed in three groups. One group received housing vouchers that could be used only in low-poverty areas (<10 percent poverty) for the first year in addition to counseling to help them find housing. After a year, they could use their vouchers anywhere. One group received vouchers that could be used anywhere but no counseling. A third control group did not receive vouchers but were eligible for any other government assistance for which they qualified. Prior studies examined the effects of the MTO intervention on labor market, educational and health outcomes.[51] The focus of the Galiani, Murphy, and Pantano (2015) study is instead on evaluating a range of counterfactual policies, such as changes in the neighborhood poverty threshold that is a condition for receiving the voucher. Their analysis sample includes 541 households in Boston, of which 165 are in the control group, 172 are in the section 8 voucher group, and 204 are in the conditional treatment experimental group.

The paper develops and estimates a model in which households choose a residential neighborhood according to their preferences for neighborhood characteristics and according to their own characteristics. They consider the choice over 585 tracts that represent different neighborhoods. The model also incorporates a moving cost that depends on distance, which varies with the household's initial residence location.

As noted in the paper, a challenge in estimating these kinds of location choice models is the potential endogeneity of rent prices, because neighborhoods may have unobserved amenities that are correlated with rent levels. The usual approach to addressing this endogeneity problem is to use instruments that come from imposing exclusion restrictions.[52] Galiani, Murphy, and Pantano (2015) show that the RCT provides another way of addressing this endogeneity problem, because it generates exogenous variation in rental prices across treatment and control groups and also within groups over time (before and after the intervention), which can be used to identify the model parameters without instruments.

In estimation, Galiani, Murphy, and Pantano (2015) use location, demographic, and rent data from the control group and from the experimental group that was subject to the low-poverty restriction.[53] For model validation purposes, they hold out the treatment group that received the unrestricted voucher. They find that the estimated model successfully replicated the mobility and neighborhood choice patterns of the held-out group. They also use the model to calculate households' willingness to pay for specific neighborhood attributes (such as the percentage of residents who are poor).

Lastly, they use the model to analyze the reasons for different take-up rates in the two treatment groups, to consider counterfactual programs, and to explore questions related to optimal program design.[54] When the estimated model is used to simulate residential choices under a range of alternative poverty thresholds ranging from 2.5 percent to 20 percent, the authors find that the program take-up rate is very sensitive to the threshold level. Adopting a less stringent poverty cut-off threshold of 20 percent generates higher take-up and leads to overall lower exposure of this set of households to poor neighborhoods, arguably improving on the existing program's design.

Migration subsidies in Bangladesh. There have been multiple field experiments in

[51] See, e.g., Kling, Liebman, and Katz (2007).

[52] See, for example, Berry, Levensohn, and Pakes (1995); Bayer, Ferriera, and McMillian (2007).

[53] In estimation, they also use census tract data and require that the location shares predicted by the model match the location shares in the census data.

[54] The program take-up rate was 63 percent for the treatment group that received the unrestricted vouchers in comparison to 55 percent for the group that was subject to the low-poverty restriction.

TABLE 5
Other Programs

Study	Group used for estimation	Out-of-sample model validation?	Evaluate counterfactual programs?
Kaboski and Townsend (2002) Model of consumption, investment, and savings in Thailand	Pre-program treatment	Yes	Yes, alternative transfer programs
Bellmare and Shearer (2018) Model of worker effort with firm gift giving in Canada	Treatment and control	No	Yes, alternative payment schemes
Paarsch and Shearer (2009) Model of worker effort in Canada	Treatment and control	No	Yes, to study effect of alternative payment schemes on firm profits
Maibom (2017) Model of how ALMP affects job search and labor market outcomes in Denmark	Treatment and control	No	Yes, alternative timing of meetings/activation program interventions
Miller, de Paula, and Valente (2020) Model of contraceptive choices in Mozambique	Pre-program treatment and control	Yes	Yes, changing subjective expectations and partner fertility and contraceptive preferences

developing countries showing that small travel subsidies generate substantial migration along with increases in income and consumption over multiple years. Lagakos, Mobarak, and Waugh (2018) argue, however, that the experimental evidence is not enough to understand whether there is a spatial mismatch of workers, namely that workers are not living in the area where they would be most productive. They also note the impact estimates are not informative about welfare effects of such programs if individuals experience disutility from rural–urban migration.

Lagakos et. al. (2018) develop a dynamic model of rural–urban migration in Bangladesh and use data from a field experiment analyzed in Bryan, Chowdhury, and Mobarak (2014) that randomly allocated subsidies to individuals living in rural areas to migrate to urban areas. In their model, households are heterogeneous in their degree of permanent productivity advantage in the urban area, and they choose to locate in either an urban region or a rural region. The model incorporates seasonal income fluctuations and stochastic income shocks. It assumes that markets are incomplete and that agents insure themselves through a buffer stock of savings.[55] Individuals face both a monetary cost of migration and a nonmonetary disutility from migration that depends on past migration experience. They can migrate permanently or temporarily.

Both treatment and control groups are used to obtain model parameter estimates, by fitting model moments to data moments deirved from the RCT. The main moments

[55] As in Bewley (1977), Aiyagari (1994), and Huggett (1996).

targeted are: (i) the increase in the seasonal migration rate resulting from the subsidy, which was 22 percent; (ii) the consumption increase for those induced to migrate, which was 30 percent; and (iii) the increase in seasonal migration one year later, after the subsidies were removed, which was nine percent.

The authors find that the consumption gains from migration observed under the RCT are not due to permanent productivity gaps between urban and rural residents, as the labor mismatch hypothesis might suggest. Rather, individuals from rural regions tend to migrate to urban areas as a form of insurance at times when they face bad shocks. The migrants are negatively selected on productivity and assets. The model estimates also reveal a high nonmonetary disutility from migration, particularly for first-time migrants. The inference from the model presents a more nuanced view about the determinants of migration decisions and the welfare benefits of the migration subsidy policy.

5.5 *Other Programs*

5.5.1 *RCT Studies*

Firm-provided wage subsidies in British Columbia. Bellemarre and Shearer (2011) analyze how increases in compensation, explained to workers as acts of kindness (gift-giving), affect workers' productivity at a tree-planting firm in British Columbia, Canada. The workers' output is observable and workers are typically compensated piece rate (per tree planted), taking into account labor market conditions and the terrain in which the planting takes place. The firm implemented a field experiment in which a random sample of workers received one of two treatments–one that provided an increase of 20–28 percent in the piece-rate wage and one that provided a base wage payment of $80 on top of the piece rate (0.20 cents per tree planted). The base wage amounted to about a 40 percent increase in the daily wage.

The authors analyze the RCT impact estimates for the two incentive designs implemented. In addition, they develop and structurally estimate a model of a worker's effort decisions given a particular gift-giving scheme. In the model, a worker's effort decision depends on two key parameters: one measuring the curvature of the effort cost function and another that measures the worker's response to monetary gifts from the firm, which they call a kindness parameter.[56] After using both the control and treatment groups to identify and estimate the model parameters, the authors use the model to calculate optimal gift-giving/piece-rate contracts.[57]

The experimental results show that the base wage gift was not profitable. On the other hand, the gift increase in the piece rate was profitable, but only when the labor market conditions otherwise led to low piece rates. The study also finds substantial heterogeneity among workers in how they respond to the firm's kindness with about half of the workers reciprocating by supplying greater effort and the other half not. The estimates indicate that reciprocity is associated with a longer tenure within the firm but the tenure effect diminishes with age. The paper finds that the piece-rate gift is most profitable for workers with strongly reciprocal preferences; profit per worker increases by as much as 14 percent for certain types of workers.

Lastly, the authors use the estimated model to study questions related to optimal contract design. In particular, they analyze the effects of composite gifts that combine a base wage and a piece-rate increase, even

[56] The modeling approach was in part inspired by Rabin's (1993) theoretical work on fairness and reciprocity.

[57] The model is estimated by a two-step nonlinear least squares procedure.

though the RCT did not include such a composite gift. They conclude that workers respond much more strongly to piece-rate gifts than to composite gifts. By analyzing the effect of differing magnitude increases in the piece-rate wage, they conclude that the firm could increase profits per worker by as much as 10 percent on average, and by up to 17 percent for workers exhibiting strongly reciprocal preferences.

Another study by Paarsch and Schearer (2009) analyzes data from the same experiment, but only from the treatment arm where the piece rate was varied. The paper explores whether observed contracts are optimal and what types of contract changes, if any, could increase firm profits. The paper develops a model where firms are choosing a contract to satisfy workers' participation constraints without assuming that the firm is maximizing profits. The piece rate is chosen to satisfy the participation constraint of the marginal worker. Workers are assumed to supply effort and to maximize their income subject to an effort cost. Model parameters are estimated by maximum likelihood using both the control group and treatment group data. The paper demonstrates that the randomized variation in the piece rate under the experiment permits identification of the elasticity of effort choice (as the piece rate is varied) under weaker assumptions.

Using the estimated model, the authors derive the firm's optimal linear contract, consisting of a base rate and a piece rate, and compare profits under the optimal contract and under the observed piece-rate contract (where the base rate was zero). The results show that the difference in profits is negligible, implying that the realized contract is close to optimal. Lastly, the paper considers the possibility of tailoring contracts to specific workers by offering different base wages to workers after their productivity types are revealed. It finds that firms could potentially increase their profits by 14 percent with a tailored wage scheme.

Active labor market programs (ALMPs) in Denmark. In many European countries, participation in so-called active labor market programs (ALMPs) is a requirement for receiving unemployment insurance (UI). ALMPs take various forms, but often it includes meetings, job search assistance and workfare/activation programs. If individuals view these arrangements as costly (e.g., a tax on their leisure), then measuring the effect of ALMPs on the duration of unemployment can overstate the benefits of such programs.[58] A large literature estimates the impacts of ALMP on employment and earnings outcomes, but very few studies explore the utility costs of such programs and the mechanisms through which the treatment effects occur.

Maibom (2017) develops and estimates a dynamic discrete choice model of job search behavior using data from a Danish RCT to more fully understand the costs and benefits of such programs. In the model, individuals search for jobs and they choose a level of search intensity.[59] If they get a job offer, then they choose whether to accept the offer. They stochastically accumulate skills while employed. Job offer rates depend on the search intensity and on the unemployment duration. Individuals also receive UI benefits that may require participation in ALMPs. Participation in ALMPs can affect utility but it can also affect job offer arrival rates.

The RCT data analyzed include 3,099 individuals (ages 22–58) living in two regions. There was a control group and a treatment group in each region. The control group

[58] Heckman, Lalonde, and Smith (1999) note that it is problematic that program impact evaluation studies value labor supply at the market wage but value time spent in the nonmarket sector at a zero wage rather than a reservation wage.

[59] The model is inspired by a model of Ferrall (2012).

was required to attend caseworker meetings every third month and to participate in a labor market activation program after 9 months of unemployment (6 months for persons under age 30) and thereafter every 26 weeks. Treatment in one region consisted of an intensified meeting schedule (every other week) and treatment in the other region consisted of earlier participation in activation. The RCT impacts showed that the employment rate was significantly higher in the treated regions with no significant effect on wages.

The job search model is estimated using both the control and treatment group data and using simulated method of moments. The estimates indicate substantial costs associated with ALMP participation. Model estimates are used to calculate the monetary compensation that would make individuals indifferent between participating in an ALMP or not and the estimates show that individuals would give up about 50 percent of the UI benefit to avoid participation. This calculation allows assessment of whether the program is a worthwhile social investment by comparing the employment gains to costs, inclusive of the nonmonetary costs borne by participants. The model estimates are also used to analyze the heterogeneity in the compensating variation in relation to future prospects and the timing of treatment. The results show that traditional cost–benefit calculations that do not take the individual utility costs into account largely overstate the gains from these types of ALMPs.

Pregnancy risk information experiment in Mozambique. An important question in developing economies is why many women do not use contraception despite reporting that they do not want to become pregnant. This phenomenon is said to lead to unwanted pregnancies and increased maternal mortality due to unsafe abortions. A study by Miller, de Paula, and Valente (2020) develops and estimates a model of a woman's contraceptive choices to understand the supply- and demand-side determinants of their decisions. The authors model the contraceptive choice as a nested logit in which there are two periods, one in which the woman decides on the contraceptive choice and then the other 12 months later, when outcomes (pregnancy, STD) are realized. The choices are between no contraception, male condoms, injections, implants, and oral contraceptives, where the hormonal methods are included in one branch of the nested logit structure. The decision problem depends on expectations of outcomes and it is assumed that a woman uses subjective probabilities about the efficacy of different methods and about any expected side effects.

The model is estimated using a sample of 584 women from Mozambique. The data include women's reported subjective beliefs as well as expressed desired fertility for both the woman and her partner. Miller, de Paula, and Valente 2020 show that the women systematically understate the risk of pregnancy and overstate the efficacy of hormonal contraceptive methods. To validate the model, the authors also carried out a randomized before–after information experiment that randomly informed a group of women about their risk of pregnancy over the next 12 months. The model is estimated using the combined treatment and control groups at baseline, prior to receiving the intervention, and it used to predict the results of the information experiment. The study finds that women who initially understate pregnancy risk and who receive the information treatment intervention increase their reported intention to use contraception by 4.4 percentage points in the experiment, which is close to the model's prediction of 4.8 percentage points.

The authors also use the estimated model to evaluate the effects of a number

TABLE 6
Studies of Programs with Spillover/GE Effects

Study	Group used for estimation	Out-of-sample model validation?	Evaluate counterfactual programs?
Kremer et al. (2011) Discrete choice model of water source in Kenya	Control and treatment	No	Yes, private versus communal property rights
Allende, Gallego, and Nielson (2011) Equil. model of school choice and program in GE and analyze effect school competition in Chile of binding capacity constraints	Control and treatment	No	Yes, extend to universal
Gautier et al. (2018) Job search model in Denmark	Control and treatment	No	Yes, universal program

of potential policy interventions. They find that supply-side interventions that increase availability or decrease costs have relatively small effects (1 percentage point reduction), in part because contraception is already widely available at low cost. However, some of the demand-side interventions they consider generate significant impacts on contraceptive use. In particular, increasing the male partner's approval of contraceptive use and aligning the male partner's desired fertility level with that of the woman increases contraceptive use by 2–4 percentage points. Also, providing women with more accurate information about pregnancy risk significantly increases contraceptive use. Another result is that women's contraceptive choices are not very sensitive to STD risk. Overall, these findings suggest that there is a potential scope for reducing fertility by providing women with accurate health information about pregnancy risk. Another implication of the results is that policy interventions that aim to influence child-bearing preferences should involve male partners.

5.5.2 *Quasi-experimental Studies*

Microfinance program in Thailand. Microfinance programs are viewed as an important mechanism for stimulating investment in developing countries. However, there are few estimates of the economic returns from such programs. Kaboski and Townsend (2011) (KT) develop and estimate a model of credit-constrained households and they use the model to compare microfinance programs to direct transfer schemes. In particular, they estimate 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 validate the model using post-program data.

The Thai Million Baht program, begun in 2001, transferred one million baht (about $25,000) to each of almost 80,000 villages in Thailand to start village banks that lend to households. KT view the program as an unanticipated exogenous quasi-experimental increase in credit. The data analysis samples come from the Townsend Thai project, which gathered panel data on rural and semi-urban households and businesses from 64 villages in four Thai provinces from 1997 to the present.

The model is based on the standard buffer stock model of savings behavior under income uncertainty (e.g., Aiyagiri 1994 and Deaton 1991). In the model, households start the first period with some level of permanent income and liquid wealth and a potential investment project of a given size. Each period, the household makes a decision about whether to undertake the investment project. The household maximizes the expected discounted value of utility over an infinite horizon. The model is estimated by generalized method of moments (GMM) using the first five years of "preexperiment" data.

The validity of the estimated model is assessed by comparing the model's predictions of the effects of the Thai Million Baht program on consumption, investment, and the probability of investing to the actual effects observed after the program was introduced. The program is incorporated into the model as a reduction in borrowing constraints by an amount that would increase the amount of total expected credit (as calculated from the model) in the village by one million baht. Impact estimates obtained using the model's simulated data are very close and, in fact, not statistically different from impact estimates obtained from regressions based on actual post-program data. One of the notable model predictions that is also borne out in the data is that the impact on consumption exceeds one million baht.

After finding support for the model's accuracy in predicting program impacts, the authors use the estimated model to compare the costs of the microfinance program to the costs of a direct transfer program that would provide the same utility benefit. They find that the cost of the microfinance program is 33 percent less, attributable to the fact that the microfinance program relaxes borrowing constraints, which the transfer program does not do.[60] The results also indicate that the largest program impact is on consumption rather than investment.[61] In summary, KT demonstrate that microfinance programs are an effective means of increasing liquidity of credit-constrained households, that they positively impact both investment and consumption, and that they are more effective than a simple transfer program.

6. *Evaluating Effects of Programs with Spillover or General Equilibrium Effects*

6.1 *RCT Studies*

Inference from RCTs can be complicated when the treatment generates spillover effects on untreated persons or when there are general equilibrium effects.[62] For example, a vaccination program could have positive spillovers for people who do not receive the vaccination. Sometimes, the issue of spillover effects is addressed by using a place-based randomization design, where randomization is performed over larger units that do not interact with each other to avoid spillovers (e.g., schools rather than students within a school). Alternatively, some studies develop models that explicitly account for the spillover effects in assessing the treatment impacts. The issue of general

[60] Even households that do not use credit can be affected by the relaxation in borrowing constraints, as it lowers their need for a buffer stock of liquidity and allows them to invest and increase consumption. Households who increase their borrowing are those who have the highest marginal valuation of liquidity, which makes the village fund program more cost effective than a simple transfer program.

[61] Additionally, KT perform a counterfactual that limits the use of credit to investment rather than consumption. The restricted policy is found to be slightly more cost effective.

[62] This violates the single unit treatment value assumption (SUTVA) commonly invoked in impact evaluations.

equilibrium effects is addressed through the explicit modeling of all market participants (for example, workers and firms) and how they interact.

Spring protection in Kenya. Kremer et al. (2011) implement an RCT to evaluate the effects of a water intervention in Kenya on outcomes related to water quality and child health. The spring protection intervention seals off the source of a naturally occurring spring and encases it in concrete so that water flows from a pipe instead of seeping from the ground, which helps to avoid contaminants from other individuals accessing the water source. In Kenya, water rights are communal and owners with a spring on their property are obliged to allow neighbors to use it without charge. This arrangement provides few private incentives to owners for investing in improvements.

The RCT randomized 184 viable unprotected springs into treatment and control groups.[63] A random selection of households that regularly used each spring was interviewed at baseline and also at follow-up rounds. Analysis of the experimental impacts showed that the intervention significantly improved water quality (as measured by E. coli contamination at the source and at the household) and also improved child health, reducing the incidence of child diarrhea by 25 percent.

As the study notes, many households access water from multiple sources and spring protection can generate spillover benefits on households in the comparison group (those initially observed to be using the unprotected control group springs). These households could decide to travel to a more distant protected water source rather than use a closer unprotected source. At baseline, 15.4 percent of comparison households get at least some of their drinking water from protected springs, but the percentage rises to 24.5 percent in follow-up rounds. To address the issue of households obtaining water from from multiple sources, the authors perform a LATE analysis, using treatment assignment as an instrument for the fraction of trips taken to obtain water from a protected source. They find that more frequent access to protected water sources significantly improves household water quality.

In addition to performing the LATE analysis, Kremer et al. 2011 develop and estimate a mixed logit random utility model of households' decisions about where to obtain water. Based on household reports on the trade-offs they face between money and walking time to collect water, the authors calculated an estimated mean annual valuation for spring protection equal to US$2.96 per household. They use the estimate to derive an implied value of $769 to avoid a statistical child death, which is substantially lower than the amounts typically used by policy makers. They interpret the estimates as evidence of a low willingness to pay for preventative health in this context. Lastly, the discrete choice model is used to simulate the welfare effects of counterfactual policies, such as giving the land owner private property rights over the spring. They find that welfare is greater with communal rights than with private property rights.

Better-informed school choice in Chile. Policy makers are often concerned that low socioeconomic status (SES) families are not investing enough in their children's human capital despite high returns to investment. One argument for why underinvestment occurs is that the parents are not well informed about their options or about the returns, raising the possibility that providing better information could lead to more efficient investment levels. Allende, Gallego, and Nielson (2019) examine the effects of

[63] The treatment was administered in multiple rounds.

an information provision RCT that targeted families of pre-K children who were soon to be entering elementary schools in Chile. The intervention consisted of a video and a personalized report card that compared different local schools. The video component included messages about the importance of selecting a high-quality school for children and the importance of schooling for labor market outcomes.

The RCT took place in 2010 in 133 preschools. There were 1,612 parents who answered the baseline and follow-up surveys. The RCT impact estimates showed that the treatment intervention shifted parents' choices toward schools with higher average test scores, higher value-added test scores, higher prices, and longer distances from home. A five year follow-up of the children using administrative test score data shows that the positive treatment effects on academic achievement are sustained.

As the authors note, it is prohibitively costly to carry out the RCT on a large scale, but it would be interesting to know the policy impacts from a large-scale adoption. One of the aims of Allende, Gallego, and Nielson's (2019) study is to understand the implications of scaling up the intervention, which they term ex ante aggregate policy evaluation. To this end, the authors develop and estimate an equilibrium model of school choice and competition among schools. The demand-side model captures how parents make trade-offs between different relevant factors, such as quality of the schools, distance, and price. The model assumes that families observe noisy signals of school characteristics, and that providing them with better information can shift the relative weights that families put on price, distance, and quality. The supply side is a model of school competition in which schools choose price and quality over the short term and also can adjust capacity over the longer term. Schools are assumed to maximize profits and a quality weighted average subject to technological constraints.[64] The authors use instruments to deal with the potential endogeneity of school characteristics, which are derived from cost variation across markets and changes in Chile's school voucher policy over time.

Using the estimated model, the paper evaluates the policy effects of an at-scale evaluation (extending the intervention to all families in the market) when schools do not react, students sort, and capacity constraints bind. It also evaluates the equilibrium effects under different assumptions on how public and private schools react and how costs change. The predicted increase in the average school quality attended by low socioeconomic families is $0.06\sigma - 0.22\sigma$. The general equilibrium policy effects are somewhat larger than the partial equilibrium effects. Also, the analysis shows that binding capacity constraints can greatly limit the policy effects.

6.2 *Quasi-experimental Studies*

Active labor market program in Denmark. Gautier et al. (2018) evaluate the effects of a Danish ALMP on labor market outcomes (earnings, employment) allowing for the possibility that the program may have negative spillover effects on nonparticipating individuals. The program was implemented as an RCT in two Danish counties and provided job search assistance to randomly chosen newly unemployed workers.[65] There were 1,814 individuals in the treatment group and 1,937 in the control group. The estimated impacts derived from the RCT show that

64 Chile has a nationwide school voucher system and more than half of children attend private schools, which can be for-profit schools.

65 All individuals who started collecting unemployment benefits between November 2005 and February 2006 participated in the experiment. Individuals born on the first to the fifteenth of the month participated in the activation program, while individuals born on the sixteenth to the thirty-first did not receive this treatment.

the program participants found jobs more quickly than nonparticipants. (See, e.g., Graversen and van Ours 2008, Rosholm 2008).

If there are negative spillover effects of the program onto untreated individuals, then the impact estimates derived from the RCT have limited policy relevance. They do not give the average effect of the program on the treated, but rather they combine positive impacts on the treated with negative impacts on the untreated. The presence of spillover effects violates the usual SUTVA that is commonly invoked in program evaluation settings. In this context, the RCT estimates cannot be used to examine the effects of a change in treatment intensity, such as the effects of a large-scale rollout of the program.[66] To get an idea of whether the program generated negative spillover effects or not, the authors perform a difference-in-difference analysis comparing the control group living in treatment counties to individuals living in similar counties where the program was not available, which showed that the controls living in treatment counties have worse labor market outcomes.

To be able to address the question of how the treatment and treatment intensity affects both participants and nonparticipants, Gautier et al. (2018) estimate the parameters of an equilibrium search model using the method of indirect inference. Their dataset combines information from the counties where the experiment took place with individuals from other comparison group counties. They argue that using data from the RCT in combination with nonexperimental data provides auxiliary moments to estimate congestion effects in the matching process and to analyze how the supply of job vacancies responds to an increase in the search intensity of program participants. The model exploits the fact that the program induces an exogenous increase in search intensity. The authors use the estimated model to understand the effects of counterfactual programs, such as one in which all newly unemployed workers receive the treatment.

7. *Conclusions*

Structural estimation is often seen as a rival approach to reduced-form analyses. This view is especially prominent in the program/policy evaluation context, with the strongest contrast being between the experimental RCT approach and the structural modeling approach. However, as illustrated by the model of section three and by the papers, the two approaches can usefully complement each other. When done well, a field experiment identifies, as cleanly as possible and under minimal assumptions, the average impact of a policy on outcomes of interest for the treated population. If a researcher is primarily interested in learning about the average program effects of an existing program on treated individuals (or on the subgroup that complies with treatment assignment, in the case of LATE), then experimental or quasi-experimental approaches may suffice.

However, policy makers often need more information than that provided by an RCT or quasi-experiment (such as RDD) to guide their decision-making at the different stages of designing, implementing and evaluating programs. For example, prior to implementation, there is the question of how to optimally design the program to achieve particular targeting and outcome objectives and to meet cost criteria. After implementation, there is interest in understanding the mechanisms generating treatment effects, in drawing inferences about how treatment effects would vary if the program were modified in some ways and/or extended to other individuals, and in predicting treatment effects

[66] Blundell, Costa Dias, and Meghir (2003) and Ferracci, Jolivet, and van den Berg (2014) found evidence for spillover effects in the context of ALMPs.

over longer terms of exposure. Lastly, there are situations where programs can generate spillover effects on control group individuals or general equilibrium effects that make it difficult to draw inferences about impacts even from an RCT. Applying structural modeling methods to RCT data greatly enhances the scope of questions that researchers can address. The use of holdout samples, usually selected to be either the treatment or control group, for purposes of out-of-sample model validation increases the credibility of estimates derived from structural models and helps to alleviate concerns about potential misspecification. Alternatively, incorporating both the treatment and control group in estimation provides additional sources of data variation useful in identifying model parameters, possibly eliminating the need for other exclusion restrictions.

Given an RCT, the researcher who adopts a structural evaluation approach must decide on whether or not to hold out one of the groups for out-of-sample validation purposes. There are a number of factors that would affect that choice. First, the researcher needs to determine whether the model parameters can be identified using data from only one of the groups or whether data from both groups are required. Second, the researcher should consider the extent to which the development and estimation of the model involve data mining. Although there is no clear metric for data mining, our own experience is that model specifications are often chosen through an intensive iterative process that checks the within-sample model fit and then adapts model features to improve the fit. We suspect that this practice is widespread, because it is difficult to foresee what type of specification will fit all the important aspects of the data. However, if a researcher were willing to commit to an initial model specification with little or no data mining, then the best practice would be to base estimation on both control and treatment samples. Analogously, researchers who decide to use a holdout sample must commit to developing the model while resisting any temptation to look at the out-of-sample fit.

In this paper, we surveyed over twenty papers that combine experimental and structural modeling approaches to program/policy evaluation. These papers span a number of fields, including labor, development, public, and urban economics. They analyze a variety of social/economic programs, including conditional cash transfer programs, welfare programs, relocation/moving subsidy programs, active labor market programs, early childhood development programs, and information interventions. As these studies illustrate, there are many different ways to fruitfully use structural modeling in conjunction with RCT data. The most critical requirement, though, is that the experiment include data beyond simple measurement of the treatment and the outcomes. The structural approach typically models agents' choice behavior subject to constraints, either in a static or dynamic context. Empirical implementation of behavioral models requires that the key variables that enter the structural components be measured.

Through this research agenda and by observing which models produce significantly more accurate forecasts, we can slowly gain a broader understanding of what types of programs can be analyzed and with what types of models. Even the failure of models to accurately reproduce experimental benchmarks is valuable information that guides future developments. The recent recognition of the value of combining structural modeling with field experiments will likely spur further applications.

References

Aiyagari, S. Rao. 1994. "Uninsured Idiosyncratic Risk and Aggregate Saving." *Quarterly Journal of Economics* 109 (3): 659–84.

Alatas, Vivi, Ririn Purnamasari, Matthew Wai-Poi, Abhijit Banerjee, Benjamin A. Olken, and Rema Hanna. 2016. "Self-Targeting: Evidence from a Field Experiment in Indonesia." *Journal of Political Economy* 124 (2): 371–427.

Allende, Claudia, Francisco Gallego, and Christopher Neilson. 2019. "Approximating the Equilibrium Effects of Informed School Choice." Princeton University Working Paper 628.

Andrew, Alison, Orazio Attanasio, Emla Fitzsimons, Sally Grantham-McGregor, Costas Meghir, and Marta Rubio-Codina. 2018. "Impacts 2 Years after a Scalable Early Childhood Development Intervention to Increase Psychosocial Stimulation in the Home: A Follow-Up of a Cluster Randomised Controlled Trial in Colombia." *PLoS Medicine* 15 (4): Article e1002556.

Andrews, Donald W. K. 1988. "Chi-Square Diagnostic Tests for Econometric Models: Theory." *Econometrica* 56 (6): 1419–53.

Andrews, Isaiah, Matthew Gentzkow, and Jesse M. Shapiro. 2020. "Transparency in Structural Research." *Journal of Business and Economic Statistics* 38 (4): 711–22.

Andrews, Isaiah, Matthew Gentzkow, and, Jesse Shapiro. 2017. "Measuring the Sensitivity of Parameter Estimates to Estimation Moments." *Quarterly Journal of Economics* 132 (4): 1553–92.

Angelucci, Manuela, and Orazio Attanasio. 2013. "The Demand for Food of Poor Urban Mexican Households: Understanding Policy Impacts Using Structural Models." *American Economic Journal: Economic Policy* 5 (1) 146–78.

Angrist, Joshua D., and Jörn-Steffen Pischke. 2009. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton, NJ: Princeton University Press.

Angrist, Joshua D., and Jörn-Steffen Pischke. 2010. "The Credibility Revolution in Empirical Economics: How Better Research Design Is Taking the Con out of Econometrics." *Journal of Economic Perspectives* 24 (2): 3–30.

Attanasio, Orazio P., Costas Meghir, and Ana Santiago. 2012. "Education Choices in Mexico: Using a Structural Model and a Randomized Experiment to Evaluate PROGRESA." *Review of Economic Studies* 79 (1): 37–66.

Attanasio, Orazio, Sarah Cattan, Emla Fitzsimons, Costas Meghir, and Marta Rubio-Codina. 2020. "Estimating the Production Function for Human Capital: Results from a Randomized Controlled Trial in Colombia." *American Economic Review* 110 (1): 48–85.

Bayer, Patrick, Fernando Ferreira, and Robert McMillan. 2007. "A Unified Framework for Measuring Preferences for Schools and Neighborhoods." *Journal of Political Economy* 115 (4): 588–638.

Behrman, Jere R., Piyali Sengupta, Petra E. Todd. 2005. "Progressing through PROGRESA: An Impact Assessment of a School Subsidy Experiment in Rural Mexico." *Economic Development and Cultural Change* 54 (1): 237–75.

Bellemare, Charles, and Bruce Shearer. 2011. "On the Relevance and Composition of Gifts within the Firm: Evidence from Field Experiments." *International Economic Review* 52 (3): 855–82.

Belzil, Christian, Julie Pernaudet, and François Poinas. 2021. "Estimating Coherency between Survey Data and Incentivized Experimental Data." IZA Discussion Paper 14594.

Berry, Steven, James Levinsohn, and Ariel Pakes. 1995. "Automobile Prices in Market Equilibrium." *Econometrica* 63 (4): 841–90.

Bewley, Truman. 1977. "The Permanent Income Hypothesis: A Theoretical Formulation." *Journal of Economic Theory* 16 (2): 252–92.

Blaug, Mark. 1980. *The Methodology of Economics: Or, How Economists Explain*. Cambridge, UK: Cambridge University Press.

Blundell, Richard, and Andrew Shephard. 2012. "Employment, Hours of Work and the Optimal Taxation of Low-Income Families." *Review of Economic Studies* 79 (2): 481–510.

Blundell, Richard, M. Costa Dias, and Costas Meghir. 2003. "The Impact of Wage Subsidies: A General Equilibrium Approach." Unpublished.

Bonhomme, Stéphane. 2020. "Discussion of 'Transparency in Structural Research' by Isaiah Andrews, Matthew Gentzkow, and Jesse Shapiro." *Journal of Business and Economic Statistics* 38 (4): 723–25.

Bourguignon, François, Francisco H. G. Ferreira, and Phillippe G. Leite. 2003. "Conditional Cash Transfers, Schooling, and Child Labor: Micro-simulating Brazil's Bolsa Escola Program." *World Bank Economic Review* 17 (2): 229–54.

Brewer, Mike, Alan Duncan, Andrew Shephard, and María José Suarez. 2006. "Did Working Families Tax Credit Work? The Impact of In-Work Support on Labour Supply in Great Britain." *Labour Economics* 13 (6): 699–720.

Bryan, Gharad, Shyamal Chowdhury, and Ahmed Mushfiq Mobarak. 2014. "Underinvestment in a Profitable Technology: The Case of Seasonal Migration in Bangladesh." *Econometrica* 82 (5): 1671–1748.

Burtless, Gary. 1987. "The Work Response to a Guaranteed Income: A Survey of Experimental Evidence." In *Lessons from the Income Maintenance Experiments*, edited by Alicia H. Munnell. Boston: Federal Reserve Bank of Boston.

Card, David, and Dean R. Hyslop. 2005. "Estimating the Effects of a Time-Limited Earnings Subsidy for Welfare-Leavers." *Econometrica* 73 (6): 1723–70.

Carneiro, Pedro, Karsten T. Hansen, and James J. Heckman. 2003. "Estimating Distributions of Treatment Effects with an Application to the Returns to Schooling and Measurement of the Effects of Uncertainty on College Choice." *International Economic Review* 44 (2): 361–422.

Chan, Mark K., and Kai Liu. 2018. "Life-Cycle and Intergenerational Effects of Childcare Reforms." *Quantitative Economics* 9 (2): 659–706.

Choi, Eleanor Jawon. 2018. "Evaluating a Structural Model of Labor Supply and Welfare Participation: Evidence from State Welfare Reform Experiments." http://dx.doi.org/10.2139/ssrn.3289937.

Creedy, John, and Guyonne Kalb. 2005. "Discrete Hours Labour Supply Modelling: Specification, Estimation and Simulation." *Journal of Economic Surveys* 19 (5): 697–734.

Cunha, Flavio, James J. Heckman, and Susanne M. Schennach. 2010. "Estimating the Technology of Cognitive and Non-cognitive Skill Formation." *Econometrica* 78 (3): 883–931.

Cunha, Flavio, Heckman, James J. and Salvador Navarro. 2007. "The Identification and Economic Content of Ordered Choice Models with Stochastic Thresholds." *International Economic Review* 48 (4): 1273–1309.

Deaton Angus. 2010. "Instruments, Randomization, and Learning about Development." *Journal of Economic Literature* 48 (2): 424–55.

Deaton, Angus. 1991. "Savings and Liquidity Constraints." *Econometrica* 59 (5): 1221–48.

Dehejia, Rajeev H., and Sadek Wahba. 1999. "Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs." *Journal of the American Statistical Association* 94 (448): 1053–62.

Dehejia, Rajeev H., and Sadek Wahba. 2002. "Propensity Score–Matching Methods for Nonexperimental Causal Studies." *Review of Economics and Statistics* 84 (1): 151–61.

DellaVigna, Stefano D. 2018. "Structural Behavioral Economics." In *Handbook of Behavioral Economics: Applications and Foundations*, Vol. 1, edited by Douglas Bernheim, Stefano DellaVigna, and David Laibson, 613–723. Amsterdam: Elsevier.

Duflo, Esther, Rema Hanna, and Stephen P. Ryan. 2012. "Incentives Work: Getting Teachers to Come to School." *American Economic Review* 102 (4): 1241–78.

Ferracci, Marc, Grégory Jolivet, and Gerard J. van den Berg. 2014. "Evidence of Treatment Spillovers within Markets." *Review of Economics and Statistics* 96 (5): 812–23.

Ferrall, Christopher. 2012. "Explaining and Forecasting Results of the Self-Sufficiency Project." *Review of Economic Studies* 79 (4): 1495–1526.

Fraker, Thomas, and Robert Moffitt. 1988. "The Effect of Food Stamps on Labor Supply: A Bivariate Selection Model." *Journal of Public Economics* 35 (1): 25–56.

Galiani, Sebastian, Alvin Murphy, and Juan Pantano. 2015. "Estimating Neighborhood Choice Models: Lessons from a Housing Assistance Experiment." *American Economic Review* 105 (11): 3385–3415.

Gautier, Pieter A., Paul Muller, Bas van der Klaauw, Michael Rosholm, and Michael Svarer. 2018. "Estimating Equilibrium Effects of Job Search Assistance." *Journal of Labor Economics* 36 (4): 1073–1125.

Glewwe, Paul, Michael Kremer, Sylvie Moulin, and Eric Zitzewitz. 2004. "Retrospective vs. Prospective Analyses of School Inputs: The Case of Flip Charts in Kenya." *Journal of Development Economics* 74 (1): 251–68.

Gong, Xiaodong, and Arthur van Soest. 2002. "Family Structure and Female Labor Supply in Mexico City." *Journal of Human Resources* 37 (1): 163–91.

Grantham-McGregor, S.M., C. A. Powell, S. P. Walker and J. H. Himes. 1991. "Nutritional Supplementation, Psychosocial Stimulation, and Mental Development of Stunted Children: The Jamaican Study." *Lancet* 338 (8758): 1–5.

Graversen, Brian K., and Jan C. van Ours. 2008. "How to Help Unemployed Find Jobs Quickly: Experimental Evidence from a Mandatory Activation Program." *Journal of Public Economics* 92 (10–11): 2020–35.

Griffen, Andrew S., and Petra E. Todd. 2017. "Assessing the Performance of Nonexperimental Estimators for Evaluating Head Start." *Journal of Labor Economics* 35 (S1): S7–63.

Hansen, Jorgen, and Xingfei Liu. 2015. "Estimating Labour Supply Responses and Welfare Participation: Using A Natural Experiment to Validate a Structural Labour Supply Model." *Canadian Journal of Economics* 48 (5): 1831–54.

Heckman, James J. 1979. "Sample Selection Bias as a Specification Error." *Econometrica* 47 (1): 153–61.

Heckman, James J. 1984. "The X^2 Goodness of Fit Statistic for Models with Parameters Estimated from Microdata." *Econometrica* 52 (6): 1543–47.

Heckman, James J. 1990. "Varieties of Selection Bias." *American Economic Review* 80 (2): 313–18.

Heckman, James J. 2001. "Micro Data, Heterogeneity and the Evaluation of Public Policy: Nobel Lecture." *Journal of Political Economy* 109 (4): 673–748.

Heckman, James J. 2010. "Building Bridges between Structural and Program Evaluation Approaches to Evaluating Policy." *Journal of Economic Literature* 48 (2): 356–98.

Heckman, James J., and V. Joseph Hotz. 1989. “Choosing among Alternative Nonexperimental Methods for Estimating the Impact of Social Programs: The Case of Manpower Training.” *Journal of the American Statistical Association* 84 (408): 862–80.

Heckman, James J., and Edward Vytlacil. 2005. “Structural Equations, Treatment Effects, and Econometric Policy Evaluation 1.” *Econometrica* 73 (3): 669–738.

Heckman, James J., Hidehiko Ichimura, and Petra E. Todd. 1997. “Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme.” *Review of Economic Studies* 64 (4): 605–54.

Heckman, James J., Hidehiko Ichimura, Jeffrey Smith, and Petra E. Todd. 1998. “Characterization Selection Bias Using Experimental Data.” *Econometrica* 66 (5): 1017–98.

Heckman, James J., Robert J. Lalonde, Jeffrey A. Smith. 1999. “The Economics and Econometrics of Active Labor Market Programs.” In *Handbook of Labor Economics*, Vol. 3A, edited byOrley C. Ashenfelter and David Card, 1865–2097. Amsterdam: North-Holland.

Heckman, James, and Sergio Urzua. 2010. “Comparing IV With Structural Models: What Simple IV Can and Cannot Identify.” *Journal of Econometrics* 156 (1): 27–37.

Heckman, James J., Sergio Urzua, and Edward Vytlacil. 2006. “Understanding instrumental variables in models with essential heterogeneity.” *Review of Economics and Statistics* 88 (3): 389–432.

Heckman, James, Seong Hyeok Moon, Rodrigo Pinto, Peter Savelyev, and Adam Yavitz. 2010. “Analyzing Social Experiments as Implemented: A Reexamination of the Evidence from the HighScope Perry Preschool Program.” *Quantitative Economics* 1 (1): 1–46.

Hoynes, Hilary Williamson. 1996. “Welfare Transfers in Two-Parent Families: Labor Supply and Welfare Participation Under AFDC-UP.” *Econometrica* 64 (2): 295–332.

Huggett, Mark. 1996. “Wealth Distribution in Life-Cycle Economies.” *Journal of Monetary Economics* 38 (3): 469–94.

Imbens, Guido W. 2010. “Better LATE than Nothing: Some Comments on Deaton (2009) and Heckman and Urzua (2009).” *Journal of Economic Literature* 48 (2): 399–423.

Imbens, Guido W., and Joshua D. Angrist. 1994. “Identification and Estimation of Local Average Treatment Effects.” *Econometrica* 62 (2): 467–75.

Kaboski Joseph P., Robert M. Townsend. 2011. “A Structural Evaluation of a Large-Scale Quasi-Experimental Microfinance Initiative.” *Econometrica* 79 (5): 1357–1406.

Keane, Michael and Robert Moffitt. 1998. “A Structural Model of Multiple Welfare Program Participation and Labor Supply.” *International Economic Review* 39 (3): 553–89.

Keane, Michael P., and Kenneth I. Wolpin. 2007. “Exploring the Usefulness of a Nonrandom Holdout Sample for Model Validation: Welfare Effects on Female Behavior.” *International Economic Review* 48 (4): 1351–78.

Keane, Michael P., and Kenneth I. Wolpin. 2010. “The Role of Labor and Marriage Markets, Preference Heterogeneity and the Welfare System on the Life Cycle Decisions of Black, Hispanic and White Women.” *International Economic Review* 51 (3): 851–92.

Keane, Michael P., Petra E. Todd, and Kenneth I. Wolpin. 2011. “The Structural Estimation of Behavioral Models: Discrete Choice Dynamic Programming Methods and Applications.” In *Handbook of Labor Economics*, Vol. 4A, edited by Orley Ashenfelter and David Card, 331–461. Amsterdam: Elsevier.

Kling, Jeffrey R., Jeffrey B. Liebman, and Lawrence F. Katz. 2007. “Experimental Analysis of Neighborhood Effects.” *Econometrica* 75 (1): 83–119.

Koopmans, Tjalling C. 1947. “Measurement without Theory.” *Review of Economics and Statistics* 29 (3): 161–72.

Kremer, Michael, Jessica Leino, Edward Miguel, and Alix Peterson Zwane. 2011. “Spring Cleaning: Rural Water Impacts, Valuation, and Property Rights Institutions.” *Quarterly Journal of Economics* 126 (1): 145–205.

Lagakos, David, Ahmed Mushfiq Mobarak, and Michael E. Waugh. 2018. “The Welfare Effects of Encouraging Rural–Urban Migration.” NBER Working Paper 24193.

LaLonde, Robert J. 1986. “Evaluating the Econometric Evaluations of Training Programs with Experimental Data.” *American Economic Review* 76 (4): 604–20.

Leamer, Edward E. 1978. *Specification Searches: Ad Hoc Inference with Nonexperimental Data*. New York: John Wiley & Sons.

Leamer Edward E. 1983. “Let’s Take the Con out of Econometrics.” *American Economic Review* 73 (1): 31–43.

Leite, Phillippe, Amber Narayan, Emmanuel Skoufias. 2015. “How Do Ex Ante Simulations Compare with Ex Post Evaluations? Evidence from the Impact of Conditional Cash Transfer Programs.” *Journal of Poverty Alleviation and International Development* 6 (1): 1–43.

Lise, Jeremy, Shannon Seitz, and Jeffrey Smith. 2015. “Evaluating Search and Matching Models Using

Experimental Data." *IZA Journal of Labor Economics* 4 (1): Article 16.

Lucas, Robert E., Jr. 1976. "Econometric Policy Evaluation: A Critique." *Carnegie–Rochester Conference Series on Public Policy* 1: 19–46

Lumsdaine Robin L., James H. Stock, David A. Wise. 1992. "Three Models of Retirement Computational Complexity versus Predictive Validity." In *Topics in the Economics of Aging*, edited by David Wise, 19–60. Chicago: University of Chicago Press.

Maibom, Jonas. 2017. "Assessing Welfare Effects of ALMPs: Combining a Structural Model and Experimental Data." https://conference.iza.org/conference_files/transatlantic_2017/maibom_j8499.pdf.

Marschak, J. 1953. "Economic Measurements for Policy and Prediction." In *Studies in Econometric Method*, edited by W. C. Hood and T. C. Koopmans, 1–26. New York: Wiley.

McFadden, Daniel, Antti Talvitie, S. Cosslett, I. Hasan, M. Johnson, F. Reid, and K. Train. 1977. *Demand Model Estimation and Validation*, Vol. 5. Berkeley, CA: The Institute of Transportation Studies, University of California.

Miller, Grant, Áureo De Paula, and Christine Valente. 2020. "Subjective Expectations and Demand for Contraception." NBER Working Paper 27271.

Moffitt, Robert. 1979. "The Labor Supply Response in the Gary Experiment." *Journal of Human Resources* 14 (4): 477–87.

Mogstad, Magne, Andreas Santos, and Alexander Torgovitsky. 2018. "Using instrumental variables for inference about policy relevant treatment parameters." *Econometrica* 86 (5): 1589–1619.

Paarsch, Harry J., and Bruce S. Shearer. 2009. "The response to incentives and contractual efficiency: Evidence from a field experiment." *European Economic Review* 53 (5): 481–94.

Pissarides Christopher A. 2000. *Equilibrium Unemployment Theory*. 2nd ed. Cambridge, MA: MIT Press.

Rabin, Matthew. 1993. "Incorporating Fairness into Game Theory and Economics." *American Economic Review* 83 (5): 1281–1302.

Rodríguez, Jorge. 2023. "Understanding the Effects of Workfare Policies on Child Human Capital." *Journal of Labor Economics* 41 (1): 39–75.

Rosenbaum, Paul, and Donald Rubin. 1983. "The Central Role of the Propensity Score in Observational Studies for Causal Effects." *Biometrika* 70 (1): 41–55.

Rosholm, Michael. 2008. "Experimental Evidence on the Nature of the Danish Employment Miracle." IZA Discussion Paper 3620.

Schady, Norbert, Maria Caridad Araujo, Ximena Peña, and Luis F. López-Calva. 2008. "Cash Transfers, Conditions, and School Enrollment in Ecuador [With Comments]." *Economía* 8 (2): 43–77.

Schorfheide, Frank, and Kenneth I. Wolpin. 2012. "On the Use of Holdout Samples for Model Selection." *American Economic Review* 102 (3): 477–81.

Schorfheide, Frank, and Kenneth I. Wolpin. 2016. "To Hold Out or Not to Hold Out?" *Research in Economics* 70 (2): 332–45

Schultz, T. Paul. 2004. "School Subsidies for the Poor: Evaluating a Mexican Progresa Poverty Strategy for Reducing Poverty." *Journal of Development Economics* 74 (1): 199–250.

Smith, Jeffrey A., and Petra E. Todd. 2005. "Does Matching Overcome Lalonde's Critique of Nonexperimental Estimators?" *Journal of Econometrics* 125 (1–2): 305–53.

Stone, Richard. 1954. "Linear Expenditure Systems and Demand Analysis: An Application to the Pattern of British Demand." *Economic Journal* 64 (255): 511–27.

Tincani, Michela, Fabian Kosse, and Enrico Miglino. 2022. "Subjective Beliefs and Inclusion Policies: Evidence from College Admissions." CEPR Discussion Paper 17177.

Todd, Petra E., and Kenneth I. Wolpin. 2006. "Assessing the Impact of a School Subsidy Program in Mexico: Using a Social Experiment to Validate a Dynamic Behavioral Model of Child Schooling and Fertility." *American Economic Review* 96 (5): 1384–1417.

Todd, Petra E., and Kenneth I. Wolpin. 2008. "*Ex ante* Evaluation of Social Programs." *Annales D'Economie Et De Statistique* 91/92: 263–91.

van Soest, Arthur. 1995. "Structural Models of Family Labor Supply: A Discrete Choice Approach." *Journal of Human Resources* 30 (1): 63–88.

Vytlacil, Edward J. 2002. "Independence, Monotonicity, and Latent Index Models: An Equivalence Result." *Econometrica* 70 (1): 331–41.

Wise, David A. 1985. "Behavioral Model versus Experimentation: The Effects of Housing Subsidies on Rent." In *Methods of Operations Research* 50, edited by P. Brucker and R. Pauly, 441–89. Konigstein: Verlag Anton Hain.

Wolpin, Kenneth I. 2013. *The Limits of Inference without Theory*. Cambridge, MA: MIT Press.

This article has been cited by:

1. Francisco J. Buera, Joseph P. Kaboski, Robert M. Townsend. 2023. From Micro to Macro Development. *Journal of Economic Literature* **61**:2, 471-503. [Abstract] [View PDF article] [PDF with links]