Evaluating a Structural Model of Labor Supply and Welfare Participation: Evidence from State Welfare Reform Experiments*

Eleanor Jawon Choi†

U.S. Bureau of Labor Statistics

April 3, 2012

Abstract

This paper assesses the ability of a structural labor supply model to predict the impacts of a welfare policy change by studying two state welfare reform experiments conducted in Minnesota and Vermont during the mid-1990s. I estimate and evaluate a static discrete choice model of labor supply and welfare participation that incorporates heterogeneity in preferences, fixed costs of work, and disutility associated with welfare take-up. Although this type of labor supply models have been commonly estimated and applied to welfare and tax policy simulations, there have been very few attempts to verify the predictive ability of the existing models. I use the experimental impacts of the welfare policy change in each state as a benchmark for the structural model's predictions. This approach is similar in spirit to LaLonde (1986). The utility parameters of the model are estimated using data from the Minnesota control group. First, based on the parameter estimates, I make predictions regarding labor supply, welfare participation, and government costs under the treatment group program in Minnesota and compare them with the observed effects of the Minnesota experiment. Next, I apply the parameter estimates to the Vermont control group and compare the predicted and observed impacts of the policy change in Vermont. The results show that the model fits the estimation sample very well, but is unable to replicate the observed treatment effects on labor supply and welfare participation outcomes of the two experiments. Consequently, the effect on net government costs is under-predicted by 31 to 93 percent in Minnesota. The prediction biases are even larger for the Vermont sample.

JEL Codes: C52, H53, I38, J22

---

*I am indebted to my adviser, David Lee, for his continued guidance and support. I also thank Orley Ashenfelter, David Card, Leandro Carvalho, Damon Clark, Hank Farber, Bo Honore, Dean Hyslop, Alan Krueger, Jae Won Lee, Lars Lefgren, Andriy Norets, Zhuan Pei, Anne Polivka, Jesse Rothstein, Mark Watson, Abigail Wozniak, Nathan Wozny, and seminar participants at Princeton, BLS, KDI, HKU, VUW, the 2010 EALE/SOLE Joint Conference, and the 2010 Econometric Society World Congress for helpful discussions and suggestions. I gratefully acknowledge the Princeton University Industrial Relations Section Graduate Fellowship and the AAUW American Dissertation Fellowship for financial support. The data used in this paper are derived from data files made available to researchers by MDRC. The author remains solely responsible for how the data have been used or interpreted. The views expressed in this paper are solely those of the author and do not necessarily reflect the views or policies of the U.S. Bureau of Labor Statistics. All errors are my own.

†Bureau of Labor Statistics, 2 Massachusetts Avenue NE, Suite 4945, Washington, DC 20212. Email: Choi.Eleanor@bls.gov. URL: https://sites.google.com/site/eleanorjchoi/.

1 Introduction

Over the past few decades, various income transfer programs with different financial incentive schemes have been implemented in the United States to achieve the three objectives: 1) reducing poverty; 2) increasing self-sufficiency within a low-income population; and 3) minimizing government costs. Since a program design that improves outcomes in one dimension often impairs them in another, policymakers have been challenged to design welfare programs that can accomplish all three goals.[1] In the context of the current economic hardship, where more people are in need but revenues are down, state governments may be pressured to modify existing programs. Thus, it becomes even more important to achieve a program design that effectively provides for low-income families without reducing their incentives to work or significantly raising government costs.

To design a welfare program that has the potential to achieve the three objectives, policymakers need to know how net income distribution will change and how much the program will cost the government due to changes in income maintenance programs. A typical economic analysis of a welfare program focuses on the potential labor supply responses to changes in financial incentives as labor supply outcomes are a self-sufficiency measure and greatly impact the effectiveness and overall costs of the program. Thus, designing an optimal welfare program requires an accurate prediction of labor supply responses to hypothetical welfare policy changes.

This paper evaluates the predictive ability of a canonical labor supply model that has been commonly estimated and widely applied to welfare and tax policy simulations in the literature, by using data from welfare policy experiments. When a policy change is studied through a randomized experiment, the experimental impact estimates of the policy change can serve as a benchmark with which a structural model's prediction results can be compared. This approach is similar in spirit to

[1] For example, Aid to Families with Dependent Children (AFDC), which was the main financial assistance program in the United States until it ended in 1996, was designed to raise the living standards of low-income families. By reducing benefits by one dollar for every dollar earned, the AFDC provided program participants with a substantial disincentive to work. As the welfare dependence of needy families and an increase in the caseloads became a serious social concern and raised government costs, the policy focus shifted to promoting work and enhancing self-sufficiency. This led to the development of the AFDC-waiver-based welfare reforms at the state level during the 1990s and subsequently to the nationwide reform via the passage of the Personal Responsibility and Work Opportunity Reconciliation Act (PRWORA) in 1996. The PRWORA replaced the AFDC with Temporary Assistance for Needy Families (TANF), which has been the major federal assistance program since 1996. Many of the new welfare programs were characterized by a lower implicit tax rate, often in combination with other features, such as work requirements or time limits on welfare duration. Since the AFDC-waivers and the TANF have allowed state governments to design and administer their own cash assistance programs, state governments have been actively devising and implementing state welfare programs including various financial incentive schemes to accomplish the three objectives. See Blank et al. (2002) for further details.

LaLonde (1986); Fraker and Maynard (1987); LaLonde and Maynard (1987); Heckman et al. (1987); Heckman and Hotz (1989); Friedlander and Robins (1995), which evaluate various non-experimental econometric methods by verifying whether the econometric methods can replicate experimental impacts of an employment and training program.[2] Although the predictions and evaluations are not based on a structural economic model in those studies, the proposed method can be applied to an evaluation of a structural model's predictive ability. The use of a randomized experiment as a tool for evaluating structural models is illustrated in recent studies by Lise et al. (2004), Todd and Wolpin (2006), and Duflo et al. (forthcoming). They apply the evaluation method to a search model, a dynamic model of fertility and child schooling, and a model of teacher performances. This paper is the first study to my knowledge that systematically assesses a structural labor supply model using randomized welfare policy experiments.

The static labor supply model considered in this paper is a typical discrete-hours labor supply model, similar to the type that appears in the literature, such as Keane and Moffitt (1998) and van Soest et al. (2002).[3] The models developed in these studies assume that an individual maximizes utility given a choice set of discrete hours of work and welfare participation. The representative features shared by these models include disutility associated with welfare take-up and the fixed costs of working positive hours. They also allow for unobserved heterogeneity in tastes for work and welfare participation, and the correlation between the two random preferences. These components are included in the labor supply model analyzed in this paper. Six specifications of the model that incorporate different combinations of the listed features are estimated and their prediction performances are evaluated.

The labor supply model is assessed by studying the effects of the Minnesota Family Investment Program (MFIP) and Vermont's Welfare Restructuring Project (WRP), operated as state welfare reform experiments during the mid-1990s. Individuals randomly assigned to the control group were offered the existing Aid to Families with Dependent Children (AFDC) characterized by the 100 percent benefit reduction rate. The treatment group members were provided with enhanced financial

[2] Propensity score matching is also evaluated in LaLonde (1986)'s method. See Dehejia and Wahba (1999, 2002); Heckman et al. (1997); Smith and Todd (2005).

[3] Earlier and more recent related papers that have developed this type of model include Fraker and Moffitt (1988), van Soest (1995), Hoynes (1996), Gong and van Soest (2002), Brewer et al. (2006) and Blundell and Shephard (forthcoming). Creedy and Kalb (2005) provide a nice survey of the large body of literature and explain typical discrete-hours labor supply models.

incentives that entailed lower benefit reduction rates, approximately 62 percent in Minnesota and 75 percent in Vermont, according to the new program rules. The Manpower Demonstration Research Corporation (MDRC) carried out the experiments and gathered data from the treatment and control group members.

The parameters of the utility function in the static labor supply model are estimated using data from the Minnesota control group. Based on these parameter estimates, I make predictions on outcomes related to labor supply, welfare participation and government costs under the MFIP incentive scheme and compare the predicted outcomes with the observed behavioral responses among the treatment group members in Minnesota. The comparison is made in terms of levels and changes of the outcomes. This method is called *within-state evaluation* throughout the paper and examines the performance of the model in forecasting the effects of a new program in a given environment.

Next, the model parameters are applied to the control group members of the Vermont experiment to predict the labor supply and welfare take-up consequences had they been facing the WRP instead of the AFDC.[4] Then, the predicted and observed impacts of the policy change in Vermont are compared. This procedure is called *cross-state evaluation* because parameters estimated using data from one state are applied to another. The cross-state evaluation procedure is an extension of the within-state evaluation method and is similar to cross-validation using two independent experiments as suggested by Levine (1989, 1993).[5] A model validated within a single randomized policy experiment may not necessarily guarantee valid predictions regarding new policies often conducted in new environments or among different populations. Thus, the cross-state evaluation provides a stronger test of the model's ability to predict the effect of a new policy in a new environment. This method is feasible only when multiple policy experiments are available.

The key assumption underlying the proposed evaluation procedures is that observable and unobservable factors are not systematically different on average across the control and treatment groups of each experiment. The explicit randomization in the experiments excludes the possibility that the

[4]The implicit tax rate of the AFDC program was the same across all states. However, the maximum benefits received by people with no earnings were slightly larger in Vermont than in Minnesota rendering the AFDC in Vermont slightly more generous than in Minnesota.

[5]Levine (1989, 1993) studied reemployment bonus experiments conducted in New Jersey and Illinois. A search model is calibrated by matching the observed and predicted treatment effects from one experiment. The treatment effect in the other experiment is predicted based on the model parameter estimates. Levine's work does not conduct within-state evaluation exploiting the experimental variation in the estimation state because the experimental variation is already used to estimate the structural parameters.

differences between the model-based predictions and the experimental estimates are derived from different populations' being exposed to different budget constraints. Therefore, the discrepancy between the predictions and experimental estimates can be attributed to true prediction errors of the model.

The findings are as follows. First, the model fits the estimation sample very well. In particular, the actual labor supply and welfare participation outcomes are accurately predicted by the specifications incorporating fixed costs of work. When fixed costs of work are omitted from the model, the probability of working part-time is overestimated.

Second, the within-state evaluation results indicate that the model shows poor out-of-sample prediction performances and fails to replicate experimental estimates of the treatment effects. The model is unable to fully account for the large increase in welfare participation due to a change in the incentive scheme from the AFDC to the MFIP. The welfare participation increase is under-predicted by 45 to 79 percent depending on the specifications of the model. None of the six specifications of the model captures the observed decrease in hours of work. The forecast results show a smaller decrease or even an increase instead. The increase in the net government costs per capita is under-predicted by 31 to 93 percent. Importantly, the fixed-cost specifications with superb within-sample fits do not perform better and sometimes generate larger prediction errors than do their non-fixed cost counterparts.

Lastly, the cross-state evaluation results are qualitatively similar to the within-sample evaluation results. The model is unsuccessful in replicating Vermont's experimental estimates of no change in welfare participation and net government costs. Five of the six specifications of the model predicted an increase in those outcomes. The impacts on the labor supply are very inaccurately predicted, especially by the specifications including fixed costs of work, and this result implies that better within-sample fits do not guarantee better cross-state evaluation performances.

The results of the evaluation exercises emphasize that it is importance to assess a model's extrapolation ability before the model is applied to policy simulations. Even a model with a very good within-sample fit can often produce large prediction errors when applied to out-of-sample forecasts. Thus, policy implications of a labor supply model, whose prediction ability is assessed only within the estimation sample, could be misleading.

The remainder of the paper is organized as follows. Section 2 explains experimental and structural approaches to predict the impacts of a welfare policy change on labor supply. Section 3 describes the static discrete choice model of labor supply and welfare participation and its empirical implementations. Section 4 provides an overview of the MFIP and WRP experiments, and discusses the data used for the analysis. Section 5 presents the estimation results for the model parameters. The within-program and cross-program evaluations of the model appear in Section 6. Section 7 concludes.

2 Experimental and Structural Approaches

Researchers have taken two approaches to predict the impacts of a welfare policy change on labor supply. One is using randomized experiments and the other is relying on a structural model.

Many of the welfare policy experiments conducted during the 1990s were designed to evaluate state welfare reforms. This approach was developed to overcome difficulties of using observational data in estimating labor supply responses. Specifically, estimates obtained from observational data are often contaminated by an omitted variable bias and unable to reveal a causal relationship between a welfare policy change and its labor supply consequences. Thus, experimental studies are mainly focused on achieving internally valid estimates of the impacts of a policy change. These experimental studies typically evaluate a particular policy change at a particular time, and the experimental estimates describe what happens when the current financial incentive scheme is replaced by another. However, the experimental estimates are unable to demonstrate what would happen if the policy change occurred among a different sample of individuals, or what would be the effects of policies that have not yet been experimented with. It is unclear how the results of a particular experiment conducted in one state may be applied in making predictions about a new incentive scheme in another state. Heckman (2000) comments on this problem mentioning that "the absence of explicit structural frameworks makes it difficult to cumulate knowledge across (experimental) studies."

To make those predictions, researchers employ a more structural approach, specifying an economic model of labor supply and welfare participation behaviors. There is an enormous body of literature on constructing and estimating labor supply models and on applying those models to out-of-sample predictions and policy simulations.[6] Those parametric models with behavioral assumptions enable researchers to predict the effects of a policy that has never been implemented. Furthermore, researchers frequently use the structural parameters or models estimated under a particular policy to predict the effects of different policies in new environments. The implicit assumption underlying this procedure is that the structural parameters do not vary based on policy or environment changes. [Heckman (2000); Heckman and Vytlacil (2005, 2007)]

It is important to examine whether the currently available labor supply models can accurately predict the effects of various welfare policy changes in different environments. Having a valid model with the best possible prediction ability is crucial in designing alternative welfare benefits and tax schedules that will have a greater impact on the labor supply at similar costs relative to the existing program. There have been very few attempts to verify the predictive ability of the existing models because there are usually no data available on the counterfactual post-intervention outcomes. This paper overcomes this problem by using experimental impacts of welfare policy changes as a benchmark for predicted outcomes based on a structural model.

3 Model and Estimation

This section describes a static labor supply model that has been commonly estimated and used for welfare and tax policy simulations. Similar models appear in previous work, including Fraker and Moffitt (1988), van Soest (1995), Hoynes (1996), Keane and Moffitt (1998), Gong and van Soest (2002), van Soest et al. (2002), Creedy and Kalb (2005), Brewer et al. (2006) and Blundell and Shephard (forthcoming). In these models, an individual solves a utility maximization problem by choosing hours of work and welfare participation status subject to a budget constraint incorporating available taxes and transfer programs. This type of models are often called discrete hours labor supply models because the models assume that individuals face a finite and discrete choice set of hours to avoid analytical and empirical difficulties created by nonlinear budget constraints with several convex and nonconvex kinks.[7] The models are often considered in a static framework because available data usually contain no information on savings behavior or human capital accumulations,

[6] Blundell and Macurdy (1999), Moffitt (2002), and Blundell et al. (2007) provide an overview of the labor supply literature.

[7] Earlier literature on labor supply often takes a continuous-hours approach. See Burtless and Hausman (1978); Hausman (1979, 1980, 1981, 1985a,b); Moffitt (1986).

which is also true for data used in this paper.[8] Static labor supply models assuming discrete hours choice have been popular in the literature because of their computational convenience compared to models assuming continuous hours choice.[9]

3.1 Utility and Budget Constraint

The utility function is assumed to take a quadratic form[10] of hours of work (h) and consumption (c) with an additive term for welfare participation (p),

$$U(h, c, p) = \beta_c c + \beta_h h + \alpha_c c^2 + \alpha_h h^2 + \alpha_{ch} ch + \phi p. \tag{1}$$

Marginal utility varies with respect to c and h through the second-order terms, and the cross-product term allows for complementarity or substitutability between c and h. The participation indicator, p, equals one when the individual is eligible for and participates in the welfare program, and ϕ represents the disutility, such as stigma or costs, associated with welfare program participation. The benefit take-up term is included to account for the observed non-participation of eligible individuals following Moffitt (1983). Utility is usually expected to increase with consumption, $\frac{\partial U}{\partial c} > 0$, decrease with hours of work, $\frac{\partial U}{\partial h} < 0$, and decrease with welfare take-up, $\frac{\partial U}{\partial p} < 0$. The model is estimated without any restrictions imposed on the utility function, and whether the signs and magnitudes of parameter estimates have reasonable economic interpretations is discussed in Section 5.

Consumption is assumed to equal the total net income determined by earnings and non-labor income because information on savings is unavailable from the data:

$$c = wh + pb(wh,\, Z) - t(wh,\, Z), \tag{2}$$

[8] In my data, key variables such as wages and hours of work are available only in cross-section data from a post random assignment survey. See Section 4 for further discussions on data and construction of the research sample.

[9] The advantages of a discrete-hours approach, compared with continuous-hours models, are as follows: 1) It is easier to incorporate complex budget constraints generated from a nonlinear tax schedule, welfare benefit formula, fixed costs of work, etc. By comparing direct utility levels associated with the given number of choices, we can avoid comparing indirect utility for each segment and the direct utility of each kink. The continuous hours method often becomes unmanageable because budget constraints usually have numerous segments with convex and non-convex kinks. 2) No explicit assumption regarding convex preferences is needed to construct a well-defined likelihood function; thus, the estimated substitution and income effects are not distorted by the pre-imposed Slutzky restrictions on the parameters. For more discussion regarding this issue, see Blundell and Macurdy (1999), Gong and van Soest (2002), and Blundell et al. (2007).

[10] The quadratic utility has been commonly used in applied work. van Soest et al. (2002) shows that the second-order model provides sufficient flexibility and yields estimation results similar to those of higher-order models.

where w is the hourly wage, $b(wh,\, Z)$ is benefits from the available welfare program, and $t(wh,\, Z)$ is tax payments. $b(wh,\, Z)$ represents the benefit formula of the MFIP for the Minnesota treatment group, the WRP and Food Stamps for the Vermont treatment group, and the AFDC and Food Stamps for the control groups in the two states.[11] $t(wh,\, Z)$ is the nonlinear tax function incorporating federal income taxes and the EITCs.[12] Both welfare benefits and tax payments depend on earnings, wh, as well as individual and household characteristics, Z, such as number of children.[13] Although the benefit and tax functions also depend on unearned income in principle, I assume that individuals have no unearned income. This is because information on unearned income is very incomplete, with many observations missing from the data. More details regarding the welfare benefit rules and tax formula can be found in Section 4 and Appendix.

An Individual is assumed to face nine points of working hours (h_j): 0, 10, 20, 30, 40, 50, 60, 70 and 80 hours per week. The continuous hours from the data are converted into the discrete categories: the weekly hours worked reported as 0, 1-15, 16-25, 26-35, 36-45, 46-55, 56-65, 66-75 and 76+ are assigned to each of the nine points, respectively. Along with the welfare participation decision, the choice set contains up to 18 work-welfare choice alternatives; (h_j, p_j), $j = 1, \ldots, J$, $J \leq 18$. Some work-welfare combinations could be infeasible if a particular choice of hours generated an income level that is too high for the individual to be eligible for the available welfare program.

An individual's utility maximization problem can be summarized as choosing a labor supply and welfare participation combination (h_j, p_j) that yields the highest utility in equation (1) subject to the budget constraint in equation (2).

3.2 Econometric Specifications and Estimation

A stochastic structure is added to enable empirical implementations of the discrete choice model of labor supply and welfare participation. The econometric specifications also include heterogeneity in preferences and fixed costs of work to make the model more realistic.

[11] Food Stamp benefits are treated as equivalent to the same amount of cash and included as a part of other cash transfer benefits. That is, participation in the Food Stamp program is not explicitly modeled, following Hoynes (1996).

[12] State taxes and the state EITCs are not incorporated as in many previous studies.

[13] Individuals with more children receive higher welfare benefits and EITCs given the same earnings level.

Alternative specific error terms are added to the utility as

$$V_j = U(h_j, c_j, p_j) + \varepsilon_j = U_j + \varepsilon_j,\ j = 1, \ldots, J \tag{3}$$

where the ε_j are i.i.d. and follow a type I extreme value distribution, and J stands for the number of feasible work-welfare choice possibilities. The alternative specific errors can be interpreted as optimization errors or unobserved job characteristics. With these basic features, the probability that person i chooses kth alternative is of the multinomial logit form

$$\begin{aligned} P_{ik} &= \Pr(V_{ik} \geq V_{ij} \text{ for } \forall j) \\ &= \frac{e^{U_{ik}}}{\sum_{j=1}^{J_i} e^{U_{ij}}}, \end{aligned} \tag{4}$$

where J_i is the number of feasible work-welfare choice alternatives that person i faces.

Fixed costs of work have been found to be an empirically important determinant of female labor supply and are commonly incorporated to expand the basic specification in equation (1).[14] Examples of these costs include commuting costs, child-care costs, and other work expenses such as uniforms and tools. Labor supply models ignoring these costs often lead to underestimation of work disincentive effects typically under-predicting the fraction of non-workers and over-predicting the fraction of part-time workers. Now, consumption c_j in equation (1) is replaced by $c_j - \eta$ if the individual works and η is estimated just like the other preference parameters. The utility function becomes

$$V_j = U(h_j, c_j - \eta \cdot 1(h_j > 0), p_j) + \varepsilon_j,\ j = 1, \ldots, J. \tag{5}$$

The utility function is allowed to vary with taste shifters by making the preference parameters dependent on individual and household characteristics. The preferences for work and welfare participation are modeled as a linear combination of covariates, such as education, age, and number and age of children. To explain the circumstances wherein individuals with the same observed characteristics behave differently, unobserved heterogeneity in tastes for work and welfare participation is also incorporated to the model. The additive random preference errors in the following equations

[14] My research sample consists of female heads of household, who are the majority of welfare recipients in the United States. Section 4 discusses how the research sample is constructed.

correspond to the linear terms for working hours and welfare take-up in the utility function.

$$\beta_h = \boldsymbol{X}_h \boldsymbol{\gamma} + u_h \tag{6}$$

$$\phi = \boldsymbol{X}_p \boldsymbol{\delta} + u_p \tag{7}$$

The signs and magnitudes of the coefficients, $\boldsymbol{\gamma}$ and $\boldsymbol{\delta}$, indicate how the taste shifters affect preferences for labor supply and welfare participation.

The random preference error terms are assumed to be independent of the alternative specific errors, ε_j, as well as all the covariates in the model, and to be normally distributed as

$$\begin{bmatrix} u_h \\ u_p \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_h^2 & \sigma_{hp} \\ \sigma_{hp} & \sigma_p^2 \end{bmatrix} \right).$$

The choice probabilities can be expressed in mixed logit forms, the integrals of standard logit probabilities over a density of random preference errors,

$$\begin{aligned} P_{ik} &= \Pr(V_{ik} \geq V_{ij} \text{ for } \forall j) \\ &= \int\int \Pr(V_{ik} \geq V_{ij} \text{ for } \forall j \,|\, u_h,\, u_p) f(u_h,\, u_p) du_h du_p, \end{aligned} \tag{8}$$

where f is the bivariate normal density of $(u_h,\, u_p)$, and the conditional probability, $\Pr(V_{ik} \geq V_{ij} \text{ for } \forall j \,|\, u_h,\, u_p)$, has the multinomial logit form in (4).

The log likelihood function is

$$LL = \sum_{i=1}^{N} \sum_{j=1}^{J_i} d_{ij} \ln P_{ij}\,, \tag{9}$$

where $d_{ij} = 1$ if person i chooses alternative j and zero otherwise, and each person faces a different number of feasible choice alternatives, J_i.

As there is no analytic expression for the probabilities in equation (8), the probabilities are approximated by the simulated means as follows:

$$\hat{P}_{ik} = \frac{1}{R} \sum_{r=1}^{R} \Pr(V_{ik} \geq V_{ij} \text{ for } \forall j \,|\, u_h^r,\, u_p^r),$$

where R is the number of draws and u_h^r, u_p^r, $r = 1, \ldots, R$, are independent draws from $f(u_h, u_p)$. Inserting the simulated probabilities into the log likelihood in equation (9) yields a simulated log likelihood,

$$SLL = \sum_{i=1}^{N} \sum_{j=1}^{J_i} d_{ij} \ln \hat{P}_{ij} \,. \tag{10}$$

The utility parameters (β_c, α_c, α_h, α_{ch}, $\boldsymbol{\gamma}'$, $\boldsymbol{\delta}'$), fixed costs of work (η), and the variance-covariance matrix of the random preference errors ($\sigma_h^2, \sigma_p^2, \sigma_{hp}$) are estimated by maximizing the simulated log likelihood. When the number of simulation draws, R, is fixed, a Simulated Maximum Likelihood (SML) estimator is inconsistent due to the simulation bias in $\ln \hat{P}_{ij}$. The asymptotic theory shows that an SML estimator is consistent and efficient, asymptotically equivalent to a Maximum Likelihood (ML) estimator when R tends to infinity faster than $\sqrt{N}$.[15] R is set to 50 for the estimation results in Section 5.[16]

Note that the parameters are identified from the cross-sectional variation across individuals in their hours of work, welfare participation choices, wages, individual and household characteristics. Wages, individual characteristics, and number of children are assumed to be exogenous to their underlying desire to work and participate in welfare. Fixed costs of work is identified by people who choose zero hours and thus do not incur costs. Because the model is estimated based on the Minnesota control group sample, identification does not rely on experimental variation.

4 Background and Data

This section describes the MFIP and WRP experiments conducted in Minnesota and Vermont, the welfare program rules implemented in control and treatment groups of each state, and the data used for the analysis.

4.1 MFIP and WRP Experiments

The MFIP and WRP were implemented as state welfare reforms in Minnesota and Vermont during the mid-1990s. Both of the programs were initiated under the AFDC waivers and implemented as pilot programs. The MFIP and WRP aimed to increase work, reduce dependence on welfare and

[15] See Train (2003) for more details on the theory and estimation of mixed logit models.
[16] Gong and van Soest (2002) find that $R = 20$ is large enough.

reduce poverty. To achieve these goals, the waiver-based reforms included enhanced financial work incentives and mandatory employment and training programs for long-term welfare recipients. They were evaluated through randomized experiments to demonstrate the effectiveness of new program designs compared with the existing AFDC. Many of these features were incorporated into Minnesota and Vermont's current TANF programs, each of which took effect in 1998 and 2001 following the enactment of the 1996 federal welfare reform law.[17]

Randomized evaluations of the MFIP and WRP were carried out by the MDRC from 1994 to 1998 and from 1994 to 2001, respectively. People in the evaluation counties of Minnesota (Vermont) between April 1994 (July 1994) and March 1996 (December 1996) were randomly assigned to one of the three program groups when they newly applied or reapplied for welfare: the AFDC group, the full MFIP (full WRP) group, or the MFIP (WRP) incentives-only group. Persons who were randomly assigned to one of the three groups received assistance if they were eligible for and participated in the available welfare programs. Individuals assigned to the AFDC group were subject to the original welfare program guidelines. Individuals in the MFIP (WRP) incentives-only group received public assistance according to the MFIP (WRP) financial incentive system but were not subject to the work requirement, whereas the full MFIP (full WRP) group members were enrolled in the full program including both elements.[18] To focus on the effect of financial incentives not confounded by the effect of other program features, I restrict the research sample to individuals in the AFDC group and the MFIP (WRP) incentives-only group.[19] Throughout the paper, the AFDC group is referred as the control group and the MFIP (WRP) incentives-only group as the treatment group.

Welfare benefits are calculated according to the following formula.[20] People without any non-welfare income receive the maximum benefits, G, and the maximum benefit amount is preserved until those individuals earn D, the fixed earnings disregard. The benefits are reduced by t dollars for every dollar increase in earned income, E, over the fixed earnings disregard, D. Thus, the basic welfare benefit formula can be expressed as $B = G - t \cdot \max(E - D, 0)$. If the income is greater

[17]For more details on state-wide waivers and the federal welfare reform, see Blank (2002) or chapter 2 of Grogger and Karoly (2005).

[18]Miller et al. (2000) and Scrivener et al. (2002) explain administrative details of the MFIP and WRP.

[19]The availability of the incentives-only group is the primary reason for using data from evaluations of the MFIP and WRP in this paper. Many other states' waiver-based reforms were also evaluated through a random assignment design, but only the full programs were available, and this makes it hard to disentangle the effect of financial incentives from other features, such as work requirements or time limits.

[20]I follow the notations in Blank et al. (2002).

than the break-even point, $G/t + D$, families are no longer eligible for program participation. The magnitude of the program parameters – the maximum benefit level (G), the benefit reduction rate (t)[21], and the fixed earnings disregard (D) – differ across programs. A program is known to be more generous with higher G and D and lower t. Individuals with more children are subject to higher maximum benefits (G) and fixed earnings disregard (D) according to each program rule.

Table 1 summarizes the key differences in the monthly welfare benefit formula that the treatment and control group members in Minnesota and Vermont were subject to. The AFDC program is well-known for its 100 percent benefit reduction rate ($t = 1$) for families on aid for more than four months (with these individuals constituting the majority of the research sample). With the near 100 percent implicit tax rate, welfare recipients have very little incentive to earn more than the fixed disregard amount because the total income remains nearly the same regardless of how much they earn above the earnings disregard. To reward work efforts, the MFIP and WRP feature lower benefit reduction rates of, 62 percent and 75 percent, respectively, and a higher level of fixed earnings disregard. Because an individual's income incorporates Food Stamps, federal income taxes, and the EITCs, as well as the AFDC, MFIP, or WRP, the actual budget constraint looks much more complicated than the simple formula.[22] Figure 1 illustrates weekly incomes for a head of single-parent household with two children when she is a treatment or control group member in a particular state. Weekly earnings, benefits, taxes, and the EITCs are in 1999 dollars. We can see that the MFIP is substantially more generous than the AFDC system, whereas the difference between the WRP and AFDC is not as dramatic. The notches on the graphs are created at the earnings level where an individual loses her eligibility for the AFDC or WRP. See Appendix, for more details on eligibility and program rules of the MFIP, WRP, AFDC, and Food Stamp programs, as well as on federal income tax and EITC algorithms,.

4.2 Data

Data are available from the MDRC public-use files collected from the evaluation of the MFIP and WRP. The MFIP and WRP public-use samples include data on 14,170 and 7,691 individuals, respec-

[21] t is also often called the implicit tax rate.

[22] The MFIP included Food Stamp benefits in cash, whereas the WRP replaced only the AFDC portion of public assistance and Food Stamp benefits were offered separately in Vermont. When combined with Food Stamps, WRP benefits are reduced at a rate of 76.5 percent and AFDC benefits at a rate of 71 percent in the first four months and 94 percent thereafter.

tively, in the three program groups. The sample members were interviewed just before they were randomly assigned to each program group and data on background characteristics were gathered through the baseline survey. Monthly welfare benefits and quarterly earnings, which were available from public assistance benefit records and unemployment insurance (UI) earnings records, were observed from a year before the random assignment[23] through the time when the evaluation of the MFIP and WRP ended in June 1998 and June 2001, respectively. A follow-up survey was administered to a random subset of the sample members in each state approximately 36 (Minnesota) and 42 (Vermont) months after the random assignment. 3,720 individuals responded to the 36-month survey in Minnesota with the 81 percent response rate and 1,989 individuals responded to the 42-month survey in Vermont with the 80 percent response rate. The survey collects information on self-reported hours worked and wage rates that are not available from the administrative sources.

The analysis sample of this paper includes female heads of single-parent households in the follow-up survey samples. I excluded people in the full MFIP and the full WRP groups to analyze the effect of financial incentives not confounded by the work requirement effect as mentioned above. The research sample is restricted to the follow-up survey samples because information on wages and hours of work is indispensable for the estimation of the labor supply model studied in this paper and is collected only through the follow-up surveys. Female-headed households make up about 95 (94) percent of the single-parent families in the AFDC and MFIP (WRP) incentives-only groups from the survey sample. Numerous papers on welfare reforms and EITC expansions focus on the effect of those programs on single mothers because single mothers constitute the majority of welfare recipients.[24] By focusing on single mothers, I make my results more comparable to the results of those previous studies. With an additional drop of observations with missing values in key individual characteristics, the finalized analysis sample contains 1,309 observations for Minnesota and 743 for Vermont. From the Minnesota sample, 753 are in the control group and 556 are in the treatment group, whereas there are 368 in the control group and 375 in the treatment group in the Vermont sample.

The unobserved wages for people who are not working are imputed using the predicted wages

[23] In Vermont, data from the administrative records are available for the period beginning two years before the random assignment.

[24] For example, see Eissa and Liebman (1996) and Meyer and Rosenbaum (2001).

from the log wage equation,

$$\ln w = \boldsymbol{X}_w \boldsymbol{\psi} + u_w, \tag{11}$$

where $\boldsymbol{X}_w$ includes age and education dummy variables, as well as the county-level unemployment rate. The parameter estimates are reported in Appendix Table 1. In principle, the prediction error should be considered when the predicted wages are plugged into a non-linear model. When the choice probabilities in equation (8) are simulated, u_w can be integrated out. A more sophisticated way to deal with the missing wage issue is to estimate wage equation (11) jointly with the labor supply model in Section 3. These procedures make the estimation procedure subsequently more complicated and I use the wage imputation approach that is not uncommon in the literature.[25]

4.3 Descriptive Analysis

Table 2 provides baseline descriptive statistics for the data used for estimation and evaluation of the labor supply model. In Table 2, the treatment group and control group means, standard deviations, differences in the means and the associated standard errors are presented.

The first twelve rows of Table 2 report statistics for demographic and education variables, which are included as taste shifters when the labor supply model in Section 3 is estimated. The next six rows include employment information.[26] The subsequent seven rows show the proportion of individuals in each hourly wage category only among those with non-missing wages.[27] The bottom block of rows reports the fraction of individuals in each county. Overall, the baseline characteristics are quite similar across the control and treatment groups in each state. Although there are a couple variables for which the difference is statistically significant at the 5 percent level, the differences are not large. A linear probability model of the treatment status for all the baseline characteristics in Table 2 except the hourly wage dummies and weekly hours of work categories[28] is estimated for

[25] See Fraker and Moffitt (1988); Hoynes (1996).

[26] The weekly hours of work data are missing for non-workers and a small fraction of workers. More specifically, only 139 and 87 observations in the control and treatment groups from Minnesota have non-missing values for the hours variables. The Vermont sample includes 83 control observations and 87 treatment observations with non-missing hours. The proportion of the observations with non-missing hours is not significantly different across the control and treatment groups at the 5 percent level in either state.

[27] The baselinehourly wages are available from 132 control group observations and 83 treatment group observations in the Minnesota sample. The proportion of the observations with non-missing wages is not significantly different across the two groups at the 5 percent level. From the Vermont sample, 71 control group observations and 75 treatment group observations have non-missing hourly wages, and the difference in the proportion across the two groups is not significant at the 5 percent level.

[28] These variables are excluded because they are non-missing only for a small subset of the sample.

each state. The null hypothesis of all coefficients equaling zero cannot be rejected at the 5 percent level based on the joint significance χ^2 test.

There is one thing to note regarding the randomization procedure in Minnesota. The Minnesota sample members are randomized within each county and new-applicant/re-applicant/ongoing-recipient status. Because of this randomization scheme, members of the MFIP incentive-only group are underrepresented among new-applicants by design. This has not yet been taken into account in the subsequent analysis. I plan to make a necessary adjustment by constructing a proper weight reflecting the random assignment ratio across those subgroups.

5 Estimation Results

This section provides utility parameter estimates for the labor supply and welfare participation model described in Section 3. The model is estimated using data from the Minnesota control group and the estimation results are obtained by ML or SML techniques. The mixed logit specifications incorporating unobserved heterogeneity in tastes for work and welfare are estimated using the SML method. Standard errors are calculated using the inverse of the negative Hessian matrix calculated from numerical second derivatives of the log likelihood function evaluated at the parameter estimate. The distribution representing unobserved heterogeneity in preferences for labor supply and welfare participation is assumed to be bivariate normal.

Table 3 and Appendix Table 2 contain the estimates of the utility function parameters and the covariance matrix of the random preference errors for the six specifications of the model. All the specifications in Table 3 allow the coefficients on hours (β_h) and welfare participation (ϕ) to be affected by observable taste shifters, as in equations (6) and (7), whereas Appendix Table 2 shows the results from the specifications without taste shifters. The six specifications are constructed by mixing and matching the different components of the model. Column (1) includes the basic multinomial logit (MNL) specification without fixed costs of work or unobserved heterogeneity. In the mixed logit (MixL) specification in column (3), random errors in tastes for work and welfare participation are assumed to have independent normal distributions. The specification in column (5) relaxes the independence assumption of the labor supply and welfare participation errors and is labeled MixL-Co. The specifications in columns (2), (4), and (6) add fixed costs of work to the MNL, MixL, and MixL-Co specifications in columns (1), (3), and (5), respectively. They are denoted as MNL-FC, MixL-FC, and MixL-Co-FC.

I present Appendix Table 2 first as it is more straightforward to interpret estimated utility parameters that do not vary with the covariates. MNL estimates indicate that consumption is a good, hours of work are a bad, and that they are substitutes. The utility costs associated with welfare participation, which do not vary across individuals, are significantly different from zero. Once the fixed costs of work component is added to the model, as in MNL-FC, the coefficient on hours of work becomes positive, which implies that utility increases with hours worked for part-time hours. The marginal utility with respect to hours worked becomes negative around full-time hours. The estimated fixed costs are around \$700 per week, a significant and very large figure compared to the findings in the labor supply literature. Adding fixed costs of work dramatically increases the value of the log likelihood function.[29] MixL and MixL-FC estimates show that there is a very high degree of unobserved heterogeneity in tastes for work but the large dispersion disappears once fixed costs of work are added to the model. The same pattern are observed from the MixL-Co and MixL-FC-Co estimates. Results from MixL-Co and MixL-FC-Co show a strong negative correlation between the random preference errors in the labor supply and welfare participation equations, which indicates that welfare participants are less likely to work than are welfare non-participants even without the financial incentive schemes of the existing welfare program. The magnitudes of the utility function parameters and fixed costs of work estimates are quite similar across specifications including fixed costs in columns (2), (4), and (6).

Table 3 shows that adding observable taste shifters to the labor supply and welfare participation equations slightly increases log likelihood values compared to those achieved by each of their no-taste-shifters counterparts. The covariates included in the labor supply equation are age categories, education categories, number of children, and an indicator variable showing whether the individual has a child under the age of 6. The welfare participation equation includes all of the listed variables but age categories. Age categories are excluded as they are jointly insignificant and are assumed to satisfy exclusion restrictions. A positive coefficient on a covariate in the labor supply equation implies that an increase in the covariate increases the marginal utility of working hours and, consequently, is associated with greater work effort. The estimates show that there is an arch-shaped

[29] This is also observed in Table 3.

age profile for work effort and that work effort increases with education. Number of children either reduces propensity to work or has no effect on work effort depending on the specifications. Having a child under age 6 is found to increase work effort, which is counter-intuitive, but the coefficient estimates are insignificant for all specifications. A covariate with a positive coefficient in the welfare participation equation reduces the utility costs of welfare participation, making individuals more likely to receive available welfare benefits. Specifically, a higher education level is associated with a lower preference for welfare, whereas having a child under age 6 makes individuals more likely to participate in the existing welfare program. Interpreting the impact of having more children on welfare participation is difficult because the coefficient estimates range from a significant negative value to a significant positive value depending on the specifications.

6 Evaluation Methods and Results

This section describes the within-state and cross-state evaluation procedures and provides the evaluation results. The prediction results are based on the parameter estimates shown in Table 3.[30] The key outcomes that are considered include labor supply at the extensive and intensive margins, welfare participation, earnings, net income, welfare benefit payments and net government costs. Net costs are the net cash assistance payments per capita made by the government after the EITCs are added to and the federal income taxes are subtracted from the welfare benefit payments. Net income is the total of earnings and net cash assistance payments. By comparing the observed and predicted impacts on the selected outcomes of a welfare policy change, one can assess the prediction power of the model along the three dimensions: 1) living standards, 2) work and self-sufficiency, and 3) government costs. As explained in Blank et al. (2002), reducing poverty, enhancing work effort among low-income populations, and minimizing costs are the conflicting three goals that policymakers try to achieve in designing welfare programs. To examine the whole effects of a welfare program, impacts on the outcomes along all three dimensions should be considered.

Here are notations that are used repeatedly throughout this section. $\hat{\boldsymbol{\theta}}_{\text{MN},0}$ denotes a vector of preference parameters estimated using data from the Minnesota control group (S = MN, T = 0). $h_{S,T}(\hat{\boldsymbol{\theta}}_{\text{MN},0}, \boldsymbol{X}_i)$ indicates a predicted outcome of a person i with characteristics, $\boldsymbol{X}_i$, based

[30] The prediction results based on the no taste shifter specifications in Appendix Table 2 are very similar to the ones described in this section and available upon request.

on the parameter estimate, $\hat{\boldsymbol{\theta}}_{\mathrm{MN},0}$, under the program rule in state S and treatment status T. $h_{S,T}(\hat{\boldsymbol{\theta}}_{\mathrm{MN},0}, \boldsymbol{X}_i)$ could indicate hours of work or any of the key outcomes described above.

The within-sample fit of the model is presented in Section 6.1. The out-of-sample validations using the Minnesota treatment group sample and the Vermont control and treatment group samples follow in the subsequent subsections. Within-state evaluation results are shown in Section 6.2, and cross-state evaluation results appear in Section 6.3. Throughout the evaluation procedures, I compare prediction performances of the six specifications, MNL, MNL-FC, MixL, MixL-FC, MixL-Co, and MixL-Co-FC, described in Section 5.

6.1 Within-Sample Fit

In this section, the average predicted outcomes, $\mathrm{E}[h_{\mathrm{MN},0}(\hat{\boldsymbol{\theta}}_{\mathrm{MN},0}, \boldsymbol{X}_i)|i \in \mathrm{MN}, 0]$, and the average actual outcomes, $\mathrm{E}[h_i|i \in \mathrm{MN}, 0]$ are compared to see how well the model fits the estimation sample ($S = \mathrm{MN}$, $T = 0$). Within-sample validation is usually a part of the model development process. Most of the papers developing labor supply models and applying them to policy simulations stop the evaluation process once they have shown that the model fits the estimation sample well. Because what this paper does to assess the model's extrapolation ability is a step forward from the previous work, it is important to verify that the model considered in this paper shows a reasonable within-sample fit before moving on to the next step, out-of-sample validation.

Table 4 compares the actual distribution and the predicted probabilities of choosing each of the eight grouped categories of working hours and welfare participation. The predicted probabilities are presented for the six specifications of the model. The four clustered categories of hours can be interpreted as no work, part-time work, full-time work and extra-long hours of work.[31] The Pearson's χ^2 statistics testing the null hypothesis that the predicted and actual distributions are the same are reported in the bottom row of the table.[32] The actual sample frequencies in the first column show that 40 percent of the Minnesota control group sample was not working, and that 46 percent was working full-time and not receiving AFDC benefits. Welfare recipients were less likely to work than people not receiving welfare benefits. Most of working welfare recipients had

[31] A comparison of the actual and fitted distributions of the data across the 18 outcome categories of the model is reported in Appendix Table 3. The model is unable to reflect that 40 working hours is the most preferred configuration among welfare non-participants. Because there are very few observations in some categories, I combined the nine hour categories into four groups.

[32] These tests do not take into account the fact that the predicted distributions are based on estimated parameters.

part-time jobs. This is because most of full-time workers were ineligible for the AFDC and because welfare participants were self-selected individuals with low preference for work. The specifications including fixed costs of work, MNL-FC, MixL-FC, and MixL-FC-Co, capture these patterns in the data very closely and the predicted proportions in the eight hours of work and welfare take-up categories are nearly identical to the actual frequencies. The specifications that do not incorporate fixed costs of work, MNL, MixL, and MixL-Co, over-predict the probability of working part-time and under-predict the fractions of non-workers and full-time workers. The null hypothesis that the predicted and actual probabilities are the same cannot be rejected at the 5 percent level with the three fixed-cost specifications. The χ^2 test results show that adding fixed costs of work to the model dramatically improves the within-sample fit. The test statistic values slightly improves by including unobserved heterogeneity in preferences regarding work and welfare or non-zero correlations between the two unobserved elements to the model.

Table 5 shows a comparison of the actual and fitted outcomes along the three dimensions of welfare program effects for the full estimation sample and among subgroups of the sample members according to their welfare participation status. The top panel shows that the probability of receiving welfare is precisely estimated by all the specifications. The no-fixed-cost specifications over-predict the probability of work by 6 (0.599 vs. 0.633 for MixL-Co) to 26 (0.599 vs. 0.756 for MNL) percent and under-predict average hours of work among workers by 9 (0.367 vs. 0.334 for MixL-Co) to 23 (0.367 vs. 0.284 for MNL) percent. The prediction biases are much larger among welfare recipients. An over-predicted fraction of workers among welfare recipients leads to the over-prediction of earnings, the under-prediction of welfare benefits, and the under-prediction of net government costs among welfare recipients as reported in panels B and C. The specifications incorporating fixed costs of work, on the other hand, fit the data very well and estimate subgroup outcomes much more precisely relative to no-fixed-cost specifications. The biases from predictions based on the three fixed-cost specifications are less than two percent for the fraction of workers, average hours of work, the fraction of welfare recipients, average welfare benefit payments, and average net income among all the members in the estimation sample. The prediction biases are slightly larger for net government costs and earnings but are still around two to four percent. Among the specifications incorporating fixed costs of work, the fully specified MixL-FC-Co provides the best within-sample fit with the lowest χ^2 statistic.

6.2 Within-State Evaluation

The within-state evaluation method uses the Minnesota treatment group sample ($S = \text{MN}$, $T = 1$) as the validation sample. First, the model's ability to predict the levels of various outcomes is assessed by comparing $\text{E}[h_{\text{MN},1}(\hat{\boldsymbol{\theta}}_{\text{MN},0}, \boldsymbol{X}_i)|i \in \text{MN}, 1]$ and $\text{E}[h_i|i \in \text{MN}, 1]$. This procedure simply represents the creation of out-of-sample forecasts.

Next, the model's predicted impacts of the policy change from the AFDC to the MFIP are compared to the experimental estimates of the treatment effects. This is similar to the LaLonde (1986)-type exercise. The predicted treatment effects and the experimental treatment effect estimates are calculated according to the following expressions.

$$\begin{aligned} \text{Predicted Treatment Effect} &= \text{E}[h_{\text{MN},1}(\hat{\boldsymbol{\theta}}_{\text{MN},0}, \boldsymbol{X}_i)|i \in \text{MN}, 0] \\ &\quad -\text{E}[h_{\text{MN},0}(\hat{\boldsymbol{\theta}}_{\text{MN},0}, \boldsymbol{X}_i)|i \in \text{MN}, 0] \qquad (12) \\ \text{Observed Treatment Effect} &= \text{E}[h_i|i \in \text{MN}, 1] - \text{E}[h_i|i \in \text{MN}, 0] \qquad (13) \end{aligned}$$

If there are systematic biases associated with predicting the levels of the average outcomes, these prediction biases may cancel out by taking a difference of the two predicted levels. Note that the parameter estimates are applied not to the Minnesota treatment group sample ($S = \text{MN}$, $T = 1$) but to the control group sample ($S = \text{MN}$, $T = 0$), when post-treatment outcomes are predicted. Here, I conduct a counterfactual policy simulation of what would have happened had individuals in the control group been subject to the MFIP instead of the AFDC. As long as the random assignment is valid, $\text{E}[h_{\text{MN},1}(\hat{\boldsymbol{\theta}}_{\text{MN},0}, \boldsymbol{X}_i)|i \in \text{MN}, 0]$ and $\text{E}[h_{\text{MN},1}(\hat{\boldsymbol{\theta}}_{\text{MN},0}, \boldsymbol{X}_i)|i \in \text{MN}, 1]$ should not be systematically different on average except for any differences due to sampling errors.[33] Researchers implement counterfactual policy simulations to predict the effects of a new policy when post-intervention data are unavailable.

Table 6 has the same layout as Table 5 but presents the actual outcomes and out-of-sample forecast results for the Minnesota treatment group. The actual outcomes in the first column show that the fraction of welfare recipients was higher among the treatment group members than the control group members in Table 5. Also, people receiving grants according to the MFIP benefit

[33] They are indeed similar, as can be seen from Table 6 and the MFIP columns of Appendix Table 4.

schedule are more likely to work and tend to work longer hours than those collecting AFDC benefits. This is due to the MFIP's generous financial incentive schemes, characterized by a lower implicit tax rate and higher income cutoff relative to the AFDC. The lower benefit reduction rate can increase labor supply among welfare recipients if the substitution effect is greater than the income effect. The higher income cutoffs make people working longer hours newly eligible for the MFIP benefits.

A key detail to notice from Panel A is that all the specifications tend to seriously under-predict the welfare participation rate by 15 (0.529 vs. 0.448 for MNL) to 51 (0.529 vs. 0.260 for MixL-Co) percent. The under-predicted welfare participation rates result in considerable under-predictions of welfare benefit payments and net government costs by percentage errors of comparable magnitude as shown in Panel C. Also, note that the fixed-cost specifications with very good within-sample fits do not necessarily produce smaller out-of-sample prediction biases than the no-fixed-cost specifications with much worse within-sample fits.

The employment rate in the MFIP sample is over-predicted by six the specifications. The over-prediction of the labor supply responses for the full sample seems to be exacerbated by the prediction biases in the welfare take-up status, given that the fixed-cost models predict labor supply outcomes reasonably well conditioned on the welfare participation status. The no-fixed-cost specifications perform much worse than the fixed-cost specifications in forecasting labor supply outcomes among welfare participants. The predicted fraction of workers among welfare recipients is 66 percent higher for the MNL specification and 82 percent lower for the MixL-Co specification than the actual employment rate, whereas the prediction biases from the fixed-cost specifications are between 1.4 and 20 percent. In contrast, among welfare non-participants, the no-fixed-cost specifications provide labor supply forecasts that are as accurate as the fixed-cost specifications. This implies that the fixed costs of work play a more important role determining labor supply among the welfare participants.

Table 7 shows predicted treatment effects and experimental impact estimates described in equations (12) and (13).[34] The first two columns report experimental estimates of the treatment effects and their robust standard errors. The third column presents the mechanically calculated treatment effects as another benchmark. They are calculated under the assumption of no changes in labor supply behavior. Ashenfelter (1983) describes this situation as the increase in welfare participation

[34]The actual and predicted average outcomes of the Minnesota sample under the AFDC and the MFIP that are used for the treatment effect calculations are reported in Appendix Table 4.

due only to mechanical effects without behavioral effects of the increase in benefits. Specifically, I assume the following: 1) the AFDC recipients would keep receiving welfare benefits under the new MFIP rules without changing their labor supply behavior; 2) the eligible non-participants under the AFDC rules would maintain their labor supply behavior and no take-up decisions; 3) the "windfall beneficiaries," who are not eligible for the AFDC but newly become eligible for the MFIP due to the MFIP's higher income cutoffs, would participate in the MFIP while holding their choice of working hours fixed; and 4) there would be no "switchers," individuals earning more than the MFIP income cutoff who would reduce their hours of work to make themselves eligible for the new welfare program. The next six columns exhibit the predicted treatment effects based on the six specifications of the behavioral model.

The experimental estimates of the treatment effects show a 22.6 percentage-point (75 percent) increase in the welfare participation rate, no significant change in extensive labor supply, and a 2.3 percentage-point (10 percent) decrease in hours of work, in Panel A. The 22.6 percentage-point rise in the welfare participation rate is mainly due to the MFIP's higher income cutoffs associated with the MFIP. This is similar to the mechanical prediction of a 24.3 percentage-point increase. In contrast to the lack of labor supply changes assumed in the mechanical predictions, the dominating income effect due to the MFIP's more generous benefit payments makes people decrease their working hours by 10 percent on average. And, this leads to a significant decrease in earnings per person by 13 percent (28 dollars per week) in Panel B. There is no significant change in average income because the decrease in earnings is compensated by an increase in welfare benefits of a similar magnitude as shown in Panel C. Net government costs increase by 50 percent mainly because of the large increase in the fraction of welfare recipients.

Many of the working MFIP recipients are the ones who would not have been eligible for the AFDC. An increase in the fraction of workers and average weekly hours of work among welfare recipients is mainly due to the MFIP's lower implicit tax rate on earnings and to the higher income cutoff rendering people who were working longer hours eligible for the new welfare program. This is also reflected in a significant and large increase in earnings of 916 percent, and a small decrease in welfare benefit payments of 13.6 percent and in net government costs of 4.3 percent, among welfare participants. The bigger mechanical estimates imply that these changes among welfare recipients are smaller than would have occurred without behavioral responses. This is because the income effect associated with higher benefits reduces hours of work and partly offsets the substitution effect.

None of the six specifications are able to precisely predict these patterns in the experimental estimates. All the specifications but MixL-Co capture the increase in welfare take-up with a more generous benefit formula, but the impact estimates are 45 (22.6 percentage points vs. 12.5 percentage points for MNL) to 79 (22.6 percentage points vs. 4.8 percentage points for MixL-FC-Co) percent lower compared to the observed treatment effect. The decrease in hours of work and earnings is not picked up by the model. The no-fixed-cost specifications predict almost no changes in working hours and earnings, and the fixed-cost specifications forecast that they will increase. Due to the prediction errors regarding the labor supply and welfare participation effects, the impact on net government costs is also seriously under-predicted. The predicted increase in net government costs is lower by 31 percent for the MNL specification and by 93 percent for the MixL-FC specification than the observed 33.6 percentage-point increase. The MixL-Co and MixL-FC-Co specifications are not even able to ascertain the observed direction of the change in net costs.

An important finding from the treatment effect prediction exercise is that the fixed-cost specifications with the superb within-sample fits do not perform better and sometimes generate larger prediction errors than do their no-fixed-cost counterparts. This point is well illustrated in Figure 2, which summarizes the relations between the biases from within-sample fits and from treatment effect predictions with respect to welfare participation, labor supply, earnings, net income, welfare benefit payments and net government costs. The percentage biases are in absolute values. In this section, I focus on the solid line with circular markers constructed from the within-state evaluations. The graphs on welfare participation and net government costs show that smaller within-sample prediction errors are associated with smaller biases in predicted treatment effects. No particular correlations between the within-sample fits and the accuracy in treatment effect predictions can be seen with respect to the employment rate, earnings, net income, and welfare benefits. The hours of work graph shows that the fixed-cost specifications with better within-sample fits perform worse than the no-fixed-cost specifications in predicting the effects of the policy change.

6.3 Cross-State Evaluation

The cross-state evaluation method is an extension of the within-state evaluation method described in the previous section. The only difference between the two procedures is that the parameters estimated based on the Minnesota control group are now applied to the control group in Vermont ($S = \text{VT}, T = 0$) for a counterfactual policy simulation of a policy change from the AFDC to the WRP. The predicted outcomes of the policy change in Vermont are compared with the experimental treatment effect estimates. The predicted and observed treatment effects are calculated according to the expressions below.

$$\begin{aligned}\text{Predicted Treatment Effect} &= \text{E}[h_{\text{VT},1}(\hat{\boldsymbol{\theta}}_{\text{MN},0}, \boldsymbol{X}_i)|i \in \text{VT}, 0] \\ &\quad -\text{E}[h_{\text{VT},0}(\hat{\boldsymbol{\theta}}_{\text{MN},0}, \boldsymbol{X}_i)|i \in \text{VT}, 0] \qquad (14) \\ \text{Observed Treatment Effect} &= \text{E}[h_i|i \in \text{VT}, 1] - \text{E}[h_i|i \in \text{VT}, 0] \qquad (15)\end{aligned}$$

The cross-state evaluation assesses the model's predictive ability more thoroughly by checking whether the model could accurately predict the effects of a new program operated in a new environment. The cross-state evaluation procedure is similar to Levine (1989, 1993)'s cross-validation using two independent experiments. Note that this evaluation procedure is feasible only when data from multiple experiments are available.

Table 8 compares the observed and predicted impacts of the policy change from the AFDC to the WRP in Vermont. The first two columns show that there are no significant changes in any of the listed outcomes.[35] This could be because of the modest change in the financial transfer schedule in Vermont. The AFDC program in Vermont was slightly more generous than the AFDC in Minnesota to begin with, and the reduction in the implicit tax rate was not as large as in Minnesota. It is visible from the budget constraint graphs in Figure 1. The mechanical model predicts a 8.7 percentage-point (24 percent) increase in the welfare take-up rate. The prediction error from the mechanical model seems generated by the assumption that there are no non-participants among the people who become newly eligible. The fraction of welfare recipients predicted based on the behavioral model that allows for welfare stigma predicts a smaller increase in the welfare participation. Nevertheless, the predicted change is in the opposite direction just like the mechanical calculation, and thus, the prediction errors are very large in percentage terms. A similar pattern of prediction biases across specifications is generated for net government costs. MixL-Co is an exception whose predicted

[35]The actual and predicted average outcomes in the Vermont control and treatment groups, from which the treatment effects are calculated, are reported in Appendix Table 5.

effects are all within one standard deviation of the experimental treatment effect estimates for the full WRP sample. However, this specification performs worse than the others in predicting subgroup outcomes. No changes in labor supply at the intensive and extensive margins are quite well predicted by the no-fixed-cost specifications with the predicted effects within one standard deviation of the experimental estimates. The fixed-cost model's predictions regarding impacts on labor supply outcomes are much less accurate.

The dashed lines with square markers in the graphs from Figure 2 show the relations between within-sample prediction biases from the estimation sample and prediction biases from the cross-evaluation procedure. Better within-sample fits do not guarantee better cross-evaluation performances, similar to the within-state evaluation results in the previous section. The graphs show either negative or no correlations between within-sample fits and cross-state evaluation biases. In terms of magnitude, the cross-state evaluations generally yield larger percentage biases than the within-state evaluations in predicting the treatment effects on welfare participation rates, net income, and net government costs.

7 Conclusion

Labor economists have been estimating static discrete choice labor supply models and applying these models to welfare policy simulations for years. Nevertheless, there have been very few attempts to verify the extrapolation ability of the commonly-used models.

This paper assesses the predictive ability of a canonical static discrete-hours labor supply model by studying two state welfare reform experiments conducted in Minnesota and Vermont during the mid-1990s. The MFIP and the WRP, the new transfer programs offered to the treatment group members in Minnesota and Vermont, feature lower implicit tax rates and higher income cutoffs compared to the existing AFDC. The utility parameters of the model are estimated using data from the Minnesota control group. Based on these parameter estimates, predictions are made regarding the effects of the policy change on labor supply, welfare participation, and net government costs in each state. The accuracy of the predictions are measured by comparing the predicted and observed effects.

The evaluation results show that the labor supply model studied in this paper fails to replicate the observed effects of the experiments in both states although the model fits the estimation sample very well. Importantly, specifications of the model with better within-sample fits often generate larger prediction errors. In previous work, researchers often use parameter estimates of a labor supply model for policy simulations after validating the model based on its within-sample fit performances. The results found in this paper indicate that the current practice in the literature could yield misleading policy implications.

References

Ashenfelter, Orley, "Determining Participation in Income-Tested Social Programs," *Journal of the American Statistical Association*, 1983, *78* (383), 517–525.

Blank, Rebecca M., "Evaluating Welfare Reform in the United States," *Journal of Economic Literature*, 2002, *40* (4), 1105–1166.

_ , David E. Card, and Philip K. Robins, "Financial Incentives for Increasing Work and Income Among Low-Income Families," in "Finding Jobs: Work and Welfare Reform," Russell Sage Foundation Publications, 2002.

Blundell, Richard and Andrew Shephard, "Employment, hours of work and the optimal taxation of low income families," *Review of Economic Studies*, forthcoming.

_ and Thomas Macurdy, "Chapter 27 Labor supply: A review of alternative approaches," in Orley C. Ashenfelter and David Card, eds., *Orley C. Ashenfelter and David Card, eds.*, Vol. 3, Part 1 of *Handbook of Labor Economics*, Elsevier, 1999, pp. 1559 – 1695.

_ , Thomas MaCurdy, and Costas Meghir, "Chapter 69 Labor Supply Models: Unobserved Heterogeneity, Nonparticipation and Dynamics," in James J. Heckman and Edward E. Leamer, eds., *James J. Heckman and Edward E. Leamer, eds.*, Vol. 6, Part 1 of *Handbook of Econometrics*, Elsevier, 2007, pp. 4667 – 4775.

Brewer, Mike, Alan Duncan, Andrew Shephard, and Maria Jose Suarez, "Did working families' tax credit work? The impact of in-work support on labour supply in Great Britain," *Labour Economics*, 2006, *13* (6), 699 – 720.

Burtless, Gary and Jerry A. Hausman, "The Effect of Taxation on Labor Supply: Evaluating the Gary Negative Income Tax Experiment," *The Journal of Political Economy*, 1978, *86* (6), 1103–1130.

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

Dehejia, Rajeev H. and Sadek Wahba, "Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs," *Journal of the American Statistical Association*, 1999, *94* (448), 1053–1062.

_ and _ , "Propensity Score-Matching Methods for Nonexperimental Causal Studies," *The Review of Economics and Statistics*, 2002, *84* (1), 151–161.

Duflo, Esther, Rema Hanna, and Stephen P. Ryan, "Incentives Work: Getting Teachers to Come to School," *The American Economic Review*, forthcoming.

Eissa, Nada and Jeffrey B. Liebman, "Labor Supply Response to the Earned Income Tax Credit," *The Quarterly Journal of Economics*, 1996, *111* (2), 605–637.

Fraker, Thomas and Rebecca Maynard, "The Adequacy of Comparison Group Designs for Evaluations of Employment-Related Programs," *The Journal of Human Resources*, 1987, *22* (2), 194–227.

_ **and Robert Moffitt**, "The effect of food stamps on labor supply : A bivariate selection model," *Journal of Public Economics*, 1988, *35* (1), 25 – 56.

Friedlander, Daniel and Philip K. Robins, "Evaluating Program Evaluations: New Evidence on Commonly Used Nonexperimental Methods," *The American Economic Review*, 1995, *85* (4), 923–937.

Gong, Xiaodong and Arthur van Soest, "Family Structure and Female Labor Supply in Mexico City," *The Journal of Human Resources*, 2002, *37* (1), 163–191.

Grogger, Jeffrey and Lynn A. Karoly, *Welfare Reform: Effects of a Decade of Change*, Harvard University Press, 2005.

Hausman, Jerry A., "The econometrics of labor supply on convex budget sets," *Economics Letters*, 1979, *3* (2), 171 – 174.

_ , "The effect of wages, taxes, and fixed costs on women's labor force participation," *Journal of Public Economics*, 1980, *14* (2), 161 – 194.

_ , "Labor supply," in H. Aaron and J. Pechman, eds., *How Taxes Affect Economic Behavior*, Washington, DC: Brookings Institution, 1981.

_ , "Chapter 4 Taxes and labor supply," in Alan J. Auerbach and Martin Feldstein, eds., *Alan J. Auerbach and Martin Feldstein, eds.*, Vol. 1 of *Handbook of Public Economics*, Elsevier, 1985, pp. 213 – 263.

_ , "The Econometrics of Nonlinear Budget Sets," *Econometrica*, 1985, *53* (6), 1255–1282.

Heckman, James J., "Causal Parameters and Policy Analysis in Economics: A Twentieth Century Retrospective," *The Quarterly Journal of Economics*, 2000, *115* (1), 45–97.

_ **and Edward J. Vytlacil**, "Chapter 70 Econometric Evaluation of Social Programs, Part I: Causal Models, Structural Models and Econometric Policy Evaluation," in James J. Heckman and Edward E. Leamer, eds., *James J. Heckman and Edward E. Leamer, eds.*, Vol. 6, Part 2 of *Handbook of Econometrics*, Elsevier, 2007, pp. 4779 – 4874.

_ **and Edward Vytlacil**, "Structural Equations, Treatment Effects, and Econometric Policy Evaluation," *Econometrica*, 2005, *73* (3), 669–738.

_ **and V. Joseph Hotz**, "Choosing Among Alternative Nonexperimental Methods for Estimating the Impact of Social Programs: The Case of Manpower Training," *Journal of the American Statistical Association*, 1989, *84* (408), 862–874.

_ **, Hidehiko Ichimura, and Petra E. Todd**, "Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme," *The Review of Economic Studies*, 1997, *64* (4), 605–654.

_ **, V. Joseph Hotz, and Marcelo Dabos**, "Do We Need Experimental Data To Evaluate the Impact of Manpower Training On Earnings?," *Evaluation Review*, 1987, *11* (4), 395–427.

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

Keane, Michael and Robert Moffitt, "A Structural Model of Multiple Welfare Program Participation and Labor Supply," *International Economic Review*, 1998, *39* (3), 553–589.

LaLonde, Robert and Rebecca Maynard, "How Precise Are Evaluations of Employment and Training Programs," *Evaluation Review*, 1987, *11* (4), 428–451.

LaLonde, Robert J., "Evaluating the Econometric Evaluations of Training Programs with Experimental Data," *The American Economic Review*, 1986, *76* (4), 604–620.

Levine, Phillip B., "Testing Search Theory with Reemployment Bonus Experiments: Cross-validation of Results from New Jersey and Illinois," Technical Report 257, Industrial Relations Section Princeton University October 1989.

_ , "Testing Search Theory with Reemployment Bonus Experiments: Cross-validation of Results from New Jersey and Illinois," *Eastern Economic Journal*, Spring 1993, *19* (2), 125–141.

Lise, Jeremy, Shannon Seitz, and Jeffrey Smith, "Equilibrium Policy Experiments and the Evaluation of Social Programs," Working Paper 10283, National Bureau of Economic Research February 2004.

Meyer, Bruce D. and Dan T. Rosenbaum, "Welfare, the Earned Income Tax Credit, and the Labor Supply of Single Mothers," *The Quarterly Journal of Economics*, 2001, *116* (3), 1063–1114.

Miller, Cynthia, Virginia Knox, Lisa A. Gennetian, Martey Dodoo, Jo Anna Hunter, and Cindy Redcross, *Reforming Welfare and Rewarding Work: Final Report on the Minnesota Family Investment Programs (Volume 1: Effects on Adults)*, Manpower Demonstration Research Corporation, September 2000.

Moffitt, Robert, "An Economic Model of Welfare Stigma," *The American Economic Review*, 1983, *73* (5), 1023–1035.

_ , "The Econometrics of Piecewise-Linear Budget Constraints: A Survey and Exposition of the Maximum Likelihood Method," *Journal of Business & Economic Statistics*, 1986, *4* (3), 317–328.

Moffitt, Robert A., "Chapter 34 Welfare programs and labor supply," in Alan J. Auerbach and Martin Feldstein, eds., *Alan J. Auerbach and Martin Feldstein, eds.*, Vol. 4 of *Handbook of Public Economics*, Elsevier, 2002, pp. 2393 – 2430.

Scrivener, Susan, Richard Hendra, Cindy Redcross, Dan Bloom, Charles Michalopoulos, and Johanna Walter, *Final Report on Vermont's Welfare Restructuring Project*, Manpower Demonstration Research Corporation, September 2002.

Smith, Jeffrey A. and Petra E. Todd, "Does matching overcome LaLonde's critique of nonexperimental estimators?," *Journal of Econometrics*, 2005, *125* (1-2), 305 – 353. Experimental and non-experimental evaluation of economic policy and models.

Todd, Petra E. and Kenneth I. Wolpin, "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," *The American Economic Review*, 2006, *96* (5), 1384–1417.

Train, Kenneth E., *Discrete Choice Methods with Simulation*, Cambridge University Press, 2003.

U. S. Department of Agriculture, "Characteristics of Food Stamp Households," Alexandria, VA: U. S. Department of Agriculture, Food and Consumer Service, Office of Analysis and Evaluation various years.

U. S. Department of the Treasury, "Your Federal Income Tax: Tax Guide for Individuals," Washington, DC: U. S. Department of the Treasury, Internal Revenue Service various years.

U. S. House of Representatives, Committee on Ways and Means, *Green Book, Background Material and Data on Programs within the Jurisdiction of the Committee on Ways and Means*, Washington, DC: Government Printing Office, various years.

Urban Institute, "Welfare Rules Database," various years.

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

_ **, Marcel Das, and Xiaodong Gong**, "A structural labour supply model with flexible preferences," *Journal of Econometrics*, 2002, *107* (1-2), 345 – 374.

Appendix: Welfare Benefit and Tax Algorithms

A AFDC

The well-known formula of the AFDC monthly benefits is expressed as:

$$B_A = \begin{cases} T - \max\{0, \frac{2}{3}(E-120) - D_C + Y\} & \text{First 4 months} \\ T - \max\{0, E-120 - D_C + Y\} & \text{After 4 months} \\ T - \max\{0, E-90 - D_C + Y\} & \text{After 12 months,} \end{cases}$$

if $E+Y < 1.85N$, and $B_A = 0$ if $E+Y > 1.85N$ or $B_A < 0$, where

- B_A: AFDC benefits
- T: payment standard, which is identical to the maximum benefit in Minnesota and Vermont
- N : need standard
- D_C: child-care expense deduction
- E : earned income
- Y : unearned income.

T and N vary by family size, state and year, and the parameter values are obtained from U. S. House of Representatives, Committee on Ways and Means (various years); Urban Institute (various years).

When calculating the AFDC benefit amounts and determining the program eligibility, I assume that:

1. Earnings are the only source of income. Welfare recipients and applicants in the sample have no unearned income or assets, and there is no income contribution from other household members. ($Y = 0$)
2. Welfare recipients do not claim child care expenses. ($D_C = 0$)

B Food Stamps

The Food Stamps monthly benefits are calculated as follows:

$$\begin{aligned} B_F &= \max\{M_F, G_F - 0.3I_N\} \\ I_N &= \max\{0, 0.8E + Y + B_A - 134 - S\} \\ S &= \min\{250, \max[0, R - 0.5\max(0, 0.8E + Y + B_A - 134)]\}, \end{aligned}$$

where

- B_F: Food Stamp benefits
- G_F: maximum Food Stamp benefit
- M_F: minimum Food Stamp benefit

- S: shelter expense deduction
- R: rent paid.

G_F and M_F vary by family size and year, but not by state since Food Stamps is a federal program. The Food Stamp parameter values come from U. S. House of Representatives, Committee on Ways and Means (various years); U. S. Department of Agriculture (various years).

C MFIP

The formula for the MFIP monthly benefits is:

$$B_M = \min\{1.2G_M - 0.62E, G_M\},$$

where

- B_M: MFIP benefits
- G_M: maximum MFIP benefit
- E: earned income.

The 38% earnings disregard is applied up to around 140% of the poverty line. G_M varies with family size and year, and the parameter values are from U. S. House of Representatives, Committee on Ways and Means (various years); Miller et al. (2000); Urban Institute (various years).

D WRP

We calculate the WRP monthly benefits as follows:

$$B_W = T - \max\{0.75(E - 150) - D_C + Y, 0\} \text{ if } E + Y < 1.85N$$

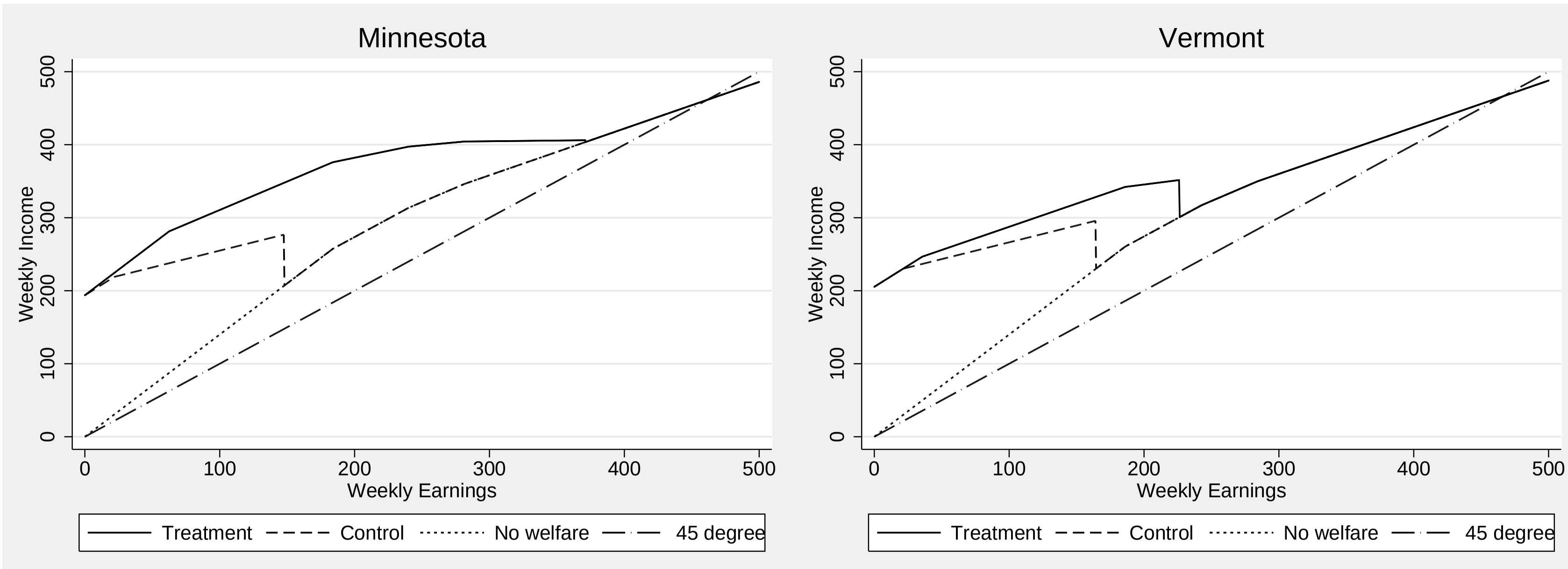
- B_W: WRP benefits
- T: payment standard, which is identical to the maximum benefit
- N : need standard
- D_C: child-care expense deduction
- E : earned income
- Y : unearned income.

T and N vary by family size and year, and the parameter values are obtained from U. S. House of Representatives, Committee on Ways and Means (various years); Scrivener et al. (2002); Urban Institute (various years). When calculating the WRP benefit amounts and determining the program eligibility, I assume $Y = 0$ and $D_C = 0$.

E Federal Income Taxes and EITCs

The federal income tax rate schedules and EITC rules are obtained from U. S. Department of the Treasury (various years).

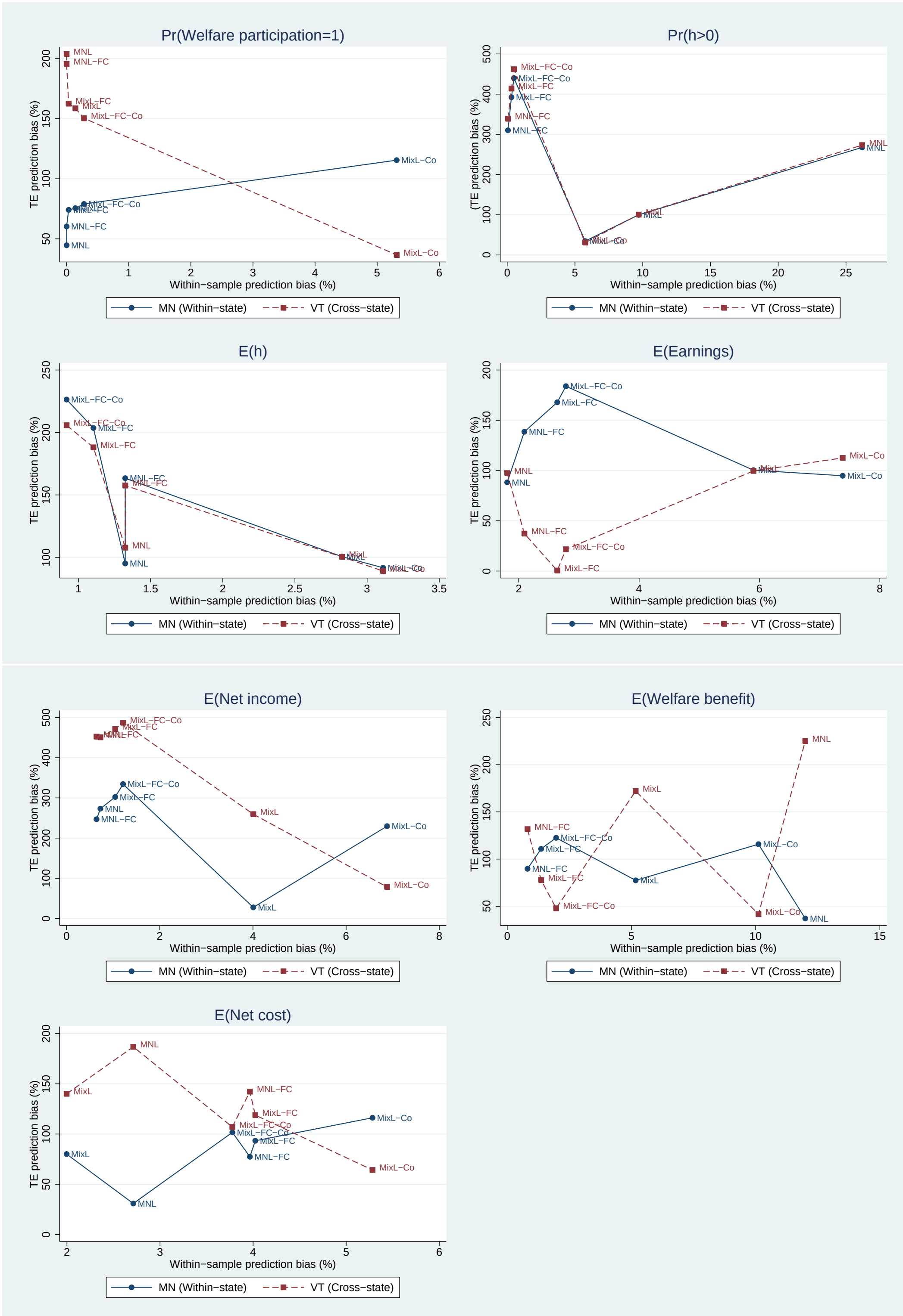
Figure 1: The Effects of Financial Incentives on Income for a Single Parent with Two Children



Notes: Weekly earnings, benefits, taxes, and the EITC are in 1999 dollars.

Sources: U. S. Department of the Treasury (various years); U. S. Department of Agriculture (various years); U. S. House of Representatives, Committee on Ways and Means (various years); Urban Institute (various years); Miller et al. (2000); Scrivener et al. (2002)

Figure 2: Biases in Predicted Treatment Effects and Within-sample Predictions



Notes: The percentage biased are all in absolute values. Within-sample prediction biases are calculated based on the results in Table 5. Treatment effect prediction biases are calculated based on the results in Table 7 and Table 8.

Table 1: Welfare Benefit Formula

Minnesota treatment group	Minnesota/Vermont control group	Vermont treatment group
MFIP	AFDC	WRP
$B_M = \min(1.2G_M - 0.62E, G_M)$	$B_A = G_A - \frac{2}{3}(E - \$120)$ if Months 1-4	$B_W = G_W - .75(E - \$150)$
Food Stamp integrated to MFIP	$G_A - (E - \$120)$ if Months 5-12	
	$G_A - (E - \$90)$ if Months > 12	
	Food Stamps	Food Stamps
	$B_F = G_F - 0.3(0.8E + B_A - \$134 - S)$	same as control group

Notes: Monthly benefits. B: benefits, G: maximum benefit, E : earned income, S: shelter expense deduction. The subscripts M, A, F, and W stand for MFIP, AFDC, Food Stamps, and WRP, respectively. Note that G_A varies by state. See Appendix for further details.
Sources: U. S. Department of the Treasury (various years); U. S. Department of Agriculture (various years); U. S. House of Representatives, Committee on Ways and Means (various years); Urban Institute (various years); Miller et al. (2000); Scrivener et al. (2002)

Table 2: Baseline Summary Statistics, by Treatment Status

	Minnesota Sample							Vermont Sample					
	MFIP		AFDC		MFIP-AFDC			WRP		AFDC		WRP-AFDC	
Variables	Mean	SD	Mean	SD	Mean Diff.	S.E.		Mean	SD	Mean	SD	Mean Diff.	S.E.
Age <20	0.049	0.215	0.077	0.267	−0.028*	0.013		0.037	0.190	0.046	0.210	−0.009	0.015
Age 20-24	0.218	0.413	0.220	0.415	−0.003	0.023		0.224	0.417	0.196	0.397	0.028	0.030
Age 25-34	0.507	0.500	0.497	0.500	0.011	0.028		0.427	0.495	0.473	0.500	−0.046	0.037
Age 35-44	0.203	0.403	0.177	0.382	0.027	0.022		0.261	0.440	0.245	0.430	0.017	0.032
Age 45+	0.023	0.151	0.029	0.169	−0.006	0.009		0.051	0.220	0.041	0.198	0.010	0.015
Number of children	2.108	1.119	1.940	1.068	0.168*	0.061		1.965	1.093	1.935	0.996	0.031	0.077
Have child under age 6	0.712	0.453	0.720	0.449	−0.008	0.025		0.568	0.496	0.579	0.494	−0.011	0.036
No degree	0.275	0.447	0.234	0.423	0.041	0.024		0.197	0.399	0.209	0.407	−0.012	0.030
GED	0.140	0.348	0.161	0.367	−0.020	0.020		0.176	0.381	0.177	0.382	−0.001	0.028
HS diploma	0.460	0.499	0.465	0.499	−0.004	0.028		0.493	0.501	0.481	0.500	0.012	0.037
Tech/AA/2 yr college	0.095	0.294	0.118	0.323	−0.023	0.017		0.088	0.284	0.098	0.297	−0.010	0.021
4 yr college	0.029	0.167	0.023	0.149	0.006	0.009		0.045	0.208	0.035	0.185	0.010	0.014
Worked FT for 6+ Mons	0.598	0.491	0.638	0.481	−0.040	0.027		0.688	0.464	0.592	0.492	0.096*	0.035
Employed	0.162	0.369	0.193	0.395	−0.031	0.021		0.205	0.404	0.193	0.395	0.012	0.029
Wkly hrs of work 1-14	0.218	0.416	0.259	0.440	−0.041	0.058		0.241	0.430	0.265	0.444	−0.024	0.067
Wkly hrs of work 15-19	0.264	0.444	0.158	0.366	0.106	0.057		0.172	0.380	0.145	0.354	0.028	0.056
Wkly hrs of work 20-29	0.253	0.437	0.295	0.458	−0.042	0.061		0.241	0.430	0.253	0.437	−0.012	0.067
Wkly hrs of work 30+	0.264	0.444	0.288	0.454	−0.023	0.061		0.345	0.478	0.337	0.476	0.007	0.073
Hrly wage $0-4	0.193	0.397	0.197	0.399	−0.004	0.056		0.240	0.430	0.310	0.466	−0.070	0.074
Hrly wage $5	0.217	0.415	0.152	0.360	0.065	0.055		0.307	0.464	0.324	0.471	−0.017	0.077
Hrly wage $6	0.181	0.387	0.288	0.454	−0.107	0.058		0.240	0.430	0.099	0.300	0.141*	0.061
Hrly wage $7	0.241	0.430	0.159	0.367	0.082	0.057		0.053	0.226	0.211	0.411	−0.158*	0.055
Hrly wage $8	0.120	0.328	0.121	0.328	−0.001	0.046		0.107	0.311	0.028	0.167	0.078	0.041
Hrly wage $9	0.024	0.154	0.030	0.172	−0.006	0.023		0.013	0.115	0.028	0.167	−0.015	0.024
Hrly wage $10-80	0.024	0.154	0.053	0.225	−0.029	0.026		0.040	0.197	0.000	0.000	0.040	0.023
County:													
Anoka	0.101	0.301	0.133	0.340	−0.032	0.018	Barre	0.149	0.357	0.152	0.360	−0.003	0.026
Dakota	0.144	0.351	0.159	0.366	−0.015	0.020	Burlington	0.323	0.468	0.312	0.464	0.010	0.034
Hennepin	0.755	0.430	0.708	0.455	0.048	0.025	Newport	0.101	0.302	0.082	0.274	0.020	0.021
							Rutland	0.176	0.381	0.201	0.401	−0.025	0.029
							Springfield	0.101	0.302	0.103	0.305	−0.002	0.022
							St. Albans	0.149	0.357	0.149	0.357	−0.000	0.026
Number of Obs	556		753					375		368			

Notes: Summary statistics are from the baseline surveys. * indicates statistical significance at the 5 percent level. The Minnesota (Vermont) sample includes 139 (83) control observations and 87 (87) treatment observations with non-missing hours of work. The information on hourly wages was available from 132 (71) control group observations and 83 (75) treatment group observation in the Minnesota (Vermont) sample.

Table 3: Results from Estimating Models of Labor Supply and Welfare Participation

	(1) MNL	(2) MNL-FC	(3) MixL	(4) MixL-FC	(5) MixL-Corr	(6) MixL-FC-Corr
Consumption/100	1.087**	0.818**	1.355	0.722**	−3.082**	0.772**
	(0.142)	(0.139)	(1.031)	(0.177)	(1.154)	(0.240)
$(\text{Consumption}/100)^2$	0.055**	0.019	−0.242	0.012	−0.721**	0.016
	(0.019)	(0.012)	(0.399)	(0.012)	(0.336)	(0.016)
Hours of work/10						
Age 20-24	−0.002	−0.000	−2.031	0.003	1.347	0.002
	(0.082)	(0.081)	(9.398)	(0.098)	(1.401)	(0.102)
Age 25-34	0.011	0.018	1.055	0.002	1.287	−0.005
	(0.080)	(0.079)	(10.501)	(0.096)	(1.343)	(0.100)
Age 35-44	0.068	0.075	2.124	0.076	3.076*	0.069
	(0.093)	(0.093)	(15.313)	(0.111)	(1.816)	(0.116)
Age 45+	−0.141	−0.113	−15.913	−0.183	−4.298	−0.206
	(0.148)	(0.140)	(18.372)	(0.172)	(2.668)	(0.172)
GED	0.058	0.024	5.652	0.022	2.201	0.014
	(0.077)	(0.083)	(13.035)	(0.084)	(1.588)	(0.088)
HS diploma	0.072	0.013	8.830	0.013	4.168**	−0.000
	(0.063)	(0.067)	(7.746)	(0.068)	(1.502)	(0.074)
Tech/AA/2 yr college	0.126	0.060	13.129	0.058	5.948**	0.038
	(0.082)	(0.090)	(8.945)	(0.091)	(2.369)	(0.099)
4 yr college	0.237	0.280	16.621	0.292	10.346**	0.288
	(0.146)	(0.186)	(11.364)	(0.193)	(5.065)	(0.199)
Number of children	0.003	0.006	−1.657	0.005	−1.052*	0.008
	(0.023)	(0.025)	(7.057)	(0.025)	(0.581)	(0.026)
Have child under age 6	0.038	0.055	0.203	0.048	1.497	0.048
	(0.054)	(0.058)	(3.759)	(0.061)	(1.622)	(0.063)
Constant	−0.792**	1.271**	9.554	1.395**	7.216**	1.424**
	(0.162)	(0.270)	(23.319)	(0.308)	(2.661)	(0.302)
$(\text{Hours of work}/10)^2$	0.113**	−0.264**	−4.018	−0.280**	−3.397**	−0.281**
	(0.022)	(0.038)	(5.449)	(0.041)	(0.942)	(0.041)
(Consumption/100) ×(Hours of work/10)	−0.248**	−0.116**	−0.109	−0.099**	1.598**	−0.109**
	(0.039)	(0.034)	(0.408)	(0.036)	(0.635)	(0.048)
Welfare participation						
GED	−0.349	−0.438	−0.481	−0.908	−2.198**	−1.177
	(0.301)	(0.338)	(0.434)	(0.846)	(1.098)	(0.975)
HS diploma	−0.675**	−0.876**	−0.960*	−1.662	−3.332**	−2.219
	(0.240)	(0.267)	(0.579)	(1.241)	(1.017)	(1.502)
Tech/AA/2 yr college	−0.937**	−1.192**	−1.379	−2.304	−5.407**	−3.113
	(0.385)	(0.417)	(0.928)	(1.807)	(1.778)	(2.184)
4 yr college	−0.319	−0.189	0.060	−1.130	−7.009**	−1.778
	(0.798)	(0.917)	(1.254)	(1.936)	(3.501)	(2.415)
Number of children	−0.267**	−0.122	0.037	−0.002	2.449**	0.040
	(0.097)	(0.108)	(0.525)	(0.228)	(0.787)	(0.265)
Have child under age 6	0.186	0.259	0.333	0.325	0.139	0.460
	(0.220)	(0.239)	(0.343)	(0.372)	(0.836)	(0.519)
Constant	−0.498	−0.201	−0.532	−0.347	−0.088	−0.578
	(0.337)	(0.380)	(0.980)	(0.622)	(1.281)	(0.877)

	(1)	(2)	(3)	(4)	(5)	(6)
Fixed costs of work/100		6.776**		7.567**		7.305**
		(1.444)		(2.174)		(2.372)
Unobserved heterogeneity						
σ_h			23.677	0.002	14.918**	0.106
			(30.683)	(0.154)	(4.234)	(0.172)
σ_p			0.551	2.253	8.587**	3.449
			(2.693)	(2.229)	(2.387)	(2.582)
σ_{hp}					−0.999**	0.612
					(0.004)	(0.242)
Log Likelihood	−1654.123	−1427.935	−1562.126	−1427.085	−1558.802	−1426.612
Number of Obs	753	753	753	753	753	753

Notes: Estimation based on data from the Minnesota control group. Number of individuals*Number of available choice alternatives = 8,538. Each individual faces up to 18 choice alternatives. Standard errors in parenthesis. * $p < 0.10$, ** $p < 0.05$.

Table 4: Actual and Predicted Probabilities for the Minnesota AFDC Sample

		Predictions					
	Actual	MNL	MNL-FC	MixL	MixL-FC	MixL-Co	MixL-FC-Co
h=0, p=0	0.130	0.068	0.125	0.109	0.126	0.137	0.127
h=0, p=1	0.271	0.176	0.276	0.234	0.273	0.229	0.271
h=10-20, p=0	0.078	0.267	0.086	0.209	0.084	0.192	0.084
h=10-20, p=1	0.027	0.124	0.022	0.066	0.024	0.055	0.025
h=30-50, p=0	0.461	0.283	0.451	0.289	0.453	0.289	0.453
h=30-50, p=1	0.005	0.002	0.005	0.003	0.005	0.002	0.006
h=60-80, p=0	0.028	0.080	0.035	0.091	0.034	0.095	0.034
h=60-80, p=1	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Pearson χ^2		353.15	2.90	198.60	1.70	184.46	1.65

Notes: Actual sample fractions and mean predicted probabilities. Predictions based on the parameter estimates reported in Table 3. Number of observations=753. $\chi^2(0.05, 7) = 14.07$.

Table 5: Actual and Predicted Outcomes for the Minnesota AFDC Sample

		Predictions					
	Actual	MNL	MNL-FC	MixL	MixL-FC	MixL-Co	MixL-FC-Co
				Panel A: Labor Supply and Welfare Participation			
All Sample							
Pr(p=1)	30.28	30.28	30.28	30.24	30.29	28.67	30.19
Pr(h>0)	59.89	75.58	59.93	65.70	60.08	63.34	60.19
E(h)	21.96	21.67	21.67	21.34	21.72	21.28	21.76
E(h|h>0)	36.67	28.41	36.01	32.11	35.99	33.37	35.99
On Welfare							
Pr(h>0|p=1)	10.53	40.31	7.56	21.68	8.57	17.35	8.99
E(h|p=1)	1.61	4.59	1.33	2.66	1.52	2.25	1.64
E(h|p=1, h>0)	15.33	10.69	12.24	11.00	12.27	11.28	12.35
Off Welfare							
Pr(h>0|p=0)	81.33	89.64	81.59	83.93	81.52	80.80	81.47
E(h|p=0)	30.80	28.67	30.15	29.03	30.19	28.64	30.17
E(h|p=0, h>0)	37.87	31.97	36.94	34.50	37.00	35.37	37.00
				Panel B: Earnings and Income			
All Sample							
E(Earnings)	212.14	208.31	207.70	199.63	206.54	196.48	206.24
E(Income)	279.01	276.99	277.22	267.83	276.10	259.82	275.63
On Welfare							
E(Earnings|p=1)	8.45	38.98	8.71	20.85	9.98	17.97	10.57
E(Income|p=1)	199.95	214.09	195.03	202.64	195.79	200.43	196.15
Off Welfare							
E(Earnings|p=0)	300.61	266.21	281.19	266.92	281.44	261.43	281.29
E(Income|p=0)	313.35	289.05	300.84	285.66	300.99	276.67	300.83
				Panel C: Cost-Benefit Outcomes			
All Sample							
E(Welfare benefits)	57.01	50.17	56.55	54.07	56.24	51.25	55.89
E(Tax payments)	10.53	9.77	9.69	9.81	9.51	10.06	9.45
E(EITC)	20.39	28.28	22.66	23.95	22.84	22.15	22.96
E(Net costs)	66.87	68.68	69.52	68.20	69.56	63.34	69.40
On Welfare							
E(Welfare benefits|p=1)	188.30	160.49	183.06	174.01	182.08	175.64	181.63
E(Tax payments|p=1)	0.00	0.00	0.00	0.00	0.00	0.00	0.00
E(EITC|p=1)	3.21	14.62	3.25	7.79	3.73	6.81	3.95
E(Net costs|p=1)	191.50	175.11	186.31	181.80	185.82	182.45	185.58
Off Welfare							
E(Tax payments|p=0)	15.11	12.12	12.34	12.83	12.36	13.24	12.35
E(EITC|p=0)	27.85	34.96	31.99	31.57	31.90	28.48	31.88
E(Net costs|p=0)	12.74	22.84	19.66	18.74	19.54	15.24	19.53

Notes: Actual and fitted outcomes. Predictions based on the parameter estimates reported in Table 3. Number of observations=753. Probabilities in percentages. Weekly hours of work. Weekly costs in 1999 dollars. Net costs = Welfare benefits - Taxes + EITCs

Table 6: Actual and Predicted Outcomes for the Minnesota MFIP Sample

		Predictions					
	Actual	MNL	MNL-FC	MixL	MixL-FC	MixL-Co	MixL-FC-Co
Panel A: Labor Supply and Welfare Participation							
All Sample							
Pr(p=1)	52.88	44.82	41.40	37.28	37.93	25.95	36.62
Pr(h>0)	57.01	79.74	65.51	64.93	68.37	61.20	69.96
E(h)	19.62	21.37	23.05	21.03	24.17	20.84	24.80
E(h|h>0)	34.52	26.64	35.08	32.00	35.26	33.76	35.38
On Welfare							
Pr(h>0|p=1)	40.82	67.61	40.23	34.72	44.44	7.53	48.21
E(h|p=1)	11.45	11.55	11.70	7.24	13.05	1.76	14.51
E(h|p=1, h>0)	28.04	17.11	27.79	20.44	28.03	22.78	28.72
Off Welfare							
Pr(h>0|p=0)	75.19	89.40	81.57	82.47	81.52	79.82	81.15
E(h|p=0)	28.79	28.75	30.21	28.74	30.25	27.30	30.05
E(h|p=0, h>0)	38.48	32.14	37.01	34.75	37.07	34.11	36.99
Panel B: Earnings and Income							
All Sample							
E(Earnings)	183.88	199.32	211.82	190.93	220.03	186.84	224.83
E(Income)	284.35	299.29	297.43	273.32	299.59	250.20	300.98
On Welfare							
E(Earnings|p=1)	85.81	100.27	95.04	59.80	106.10	16.27	117.71
E(Income|p=1)	269.08	290.19	268.84	246.56	276.96	208.01	284.66
Off Welfare							
E(Earnings|p=0)	293.93	260.98	273.85	257.67	274.18	242.47	272.62
E(Income|p=0)	301.47	285.84	297.32	277.28	297.55	260.51	295.93
Panel C: Cost-Benefit Outcomes							
All Sample							
E(Welfare benefits)	86.04	75.55	65.37	65.99	57.97	49.75	53.75
E(Tax payments)	7.86	8.50	8.41	8.71	8.48	8.91	8.54
E(EITC)	22.29	32.92	28.65	25.11	30.08	22.51	30.94
E(Net costs)	100.47	99.96	85.61	82.40	79.56	63.35	76.15
On Welfare							
E(Welfare benefits|p=1)	162.71	159.66	152.76	171.72	147.61	188.59	141.88
E(Tax payments|p=1)	0.61	0.34	0.78	0.45	0.89	0.17	1.05
E(EITC|p=1)	21.18	30.60	21.82	15.49	24.13	3.32	26.11
E(Net costs|p=1)	183.28	189.92	173.80	186.76	170.86	191.74	166.95
Off Welfare							
E(Tax payments|p=0)	15.99	11.77	11.19	12.19	11.18	11.48	11.13
E(EITC|p=0)	23.53	36.63	34.65	31.80	34.55	29.52	34.44
E(Net costs|p=0)	7.54	24.86	23.46	19.61	23.37	18.04	23.31

Notes: Actual outcomes and out-of-sample forecast results. Predictions based on the parameter estimates reported in Table 3. Number of observations=556. Probabilities in percentages. Weekly hours of work. Weekly costs in 1999 dollars. Net costs = Welfare benefits - Taxes + EITCs

Table 7: Observed and Predicted Treatment Effects (TE) of the Minnesota MFIP on Various Outcomes

	Observed		Mechanical	Predicted TE					
	TE		TE	MNL	MNL-FC	MixL	MixL-FC	MixL-Co	MixL-FC-Co
Panel A: Labor Supply and Welfare Participation Outcomes									
All Sample									
Pr(p=1)	22.6	(2.7)	24.3	12.5	9.0	5.5	5.9	−3.5	4.8
Pr(h>0)	−2.9	(2.8)	0.0	4.8	6.1	0.0	8.4	−1.9	9.8
E(h)	−2.3	(1.1)	0.0	−0.1	1.5	0.0	2.4	−0.2	3.0
E(h|h>0)	−2.2	(0.9)	0.0	−1.7	−1.0	0.0	−0.8	0.7	−0.7
On Welfare									
Pr(h>0|p=1)	30.3	(3.5)	39.8	27.5	31.1	12.8	34.1	−10.0	37.3
E(h|p=1)	9.8	(1.0)	15.7	6.8	9.6	4.3	10.6	−0.6	11.9
E(h|p=1, h>0)	12.7	(2.2)	19.1	6.1	14.5	8.9	14.7	10.8	15.3
Off Welfare									
Pr(h>0|p=0)	−6.1	(3.2)	−10.0	0.0	0.0	−1.5	0.0	−1.3	−0.3
E(h|p=0)	−2.0	(1.4)	−3.3	0.0	0.0	−0.2	0.0	−1.2	−0.2
E(h|p=0, h>0)	0.6	(1.0)	0.7	0.0	0.0	0.4	0.0	−1.0	−0.1
Panel B: Earnings and Income									
All Sample									
E(Earnings)	−28.3	(12.0)	0.0	−3.3	10.9	0.1	19.2	−1.5	23.7
E(Income)	5.3	(9.4)	16.7	19.9	18.5	6.8	21.5	−6.9	23.2
On Welfare									
E(Earnings|p=1)	77.4	(7.3)	125.7	62.5	81.6	38.0	90.7	−2.5	101.0
E(Income|p=1)	69.1	(6.8)	93.3	69.8	63.4	36.7	70.1	0.9	77.0
Off Welfare									
E(Earnings|p=0)	−6.7	(16.9)	5.3	0.0	0.0	−0.6	0.0	−10.1	−1.3
E(Income|p=0)	−11.9	(15.6)	−14.6	0.0	0.0	−3.0	0.0	−10.6	−1.3
Panel C: Cost-Benefit Outcomes									
All Sample									
E(Welfare benefits)	29.0	(5.2)	16.7	18.3	3.0	6.5	−3.2	−4.6	−6.5
E(Tax payments)	−2.7	(1.1)	0.0	−1.2	−0.5	0.0	−0.2	0.0	−0.1
E(EITC)	1.9	(1.4)	0.0	3.7	4.1	0.1	5.2	−0.8	5.9
E(Net costs)	33.6	(5.4)	16.7	23.2	7.6	6.7	2.3	−5.5	−0.6
On Welfare									
E(Welfare benefits|p=1)	−25.6	(4.9)	−53.2	−8.2	−34.5	−8.3	−38.4	7.2	−43.3
E(Tax payments|p=1)	0.6	(0.1)	1.6	0.3	0.8	0.4	0.9	0.2	1.0
E(EITC|p=1)	18.0	(1.7)	22.3	15.9	17.1	7.3	18.7	−3.6	20.4
E(Net costs|p=1)	−8.2	(4.3)	−32.4	7.3	−18.2	−1.3	−20.5	3.4	−24.0
Off Welfare									
E(Tax payments|p=0)	0.9	(1.8)	6.2	0.0	0.0	0.7	0.0	−0.3	−0.1
E(EITC|p=0)	−4.3	(1.8)	−13.7	0.0	0.0	−1.7	0.0	−0.9	−0.1
E(Net costs|p=0)	−5.2	(3.0)	−19.9	0.0	0.0	−2.4	0.0	−0.5	0.0

Notes: Experimental treatment effect estimates and within-state evaluation results. Robust standard errors of experimental estimates are in parentheses. Predictions based on the parameter estimates reported in Table 3. Net costs = Welfare benefits - Taxes + EITCs. Probability differences in percentage points. Weekly hours of work. Weekly costs in 1999 dollars. The mechanically calculated treatment effects are calculated under the assumptions: 1) the AFDC recipients would keep receiving welfare benefits under the new MFIP rules without changing their labor supply behavior; 2) the eligible non-participants under the AFDC rules would maintain their labor supply behavior and no take-up decisions; 3) the "windfall beneficiaries," who are not eligible for the AFDC but newly become eligible for the MFIP due to the MFIP's higher income cutoffs, would participate in the MFIP while holding their choice of working hours fixed; and 4) there would be no "switchers," individuals earning more than the MFIP income cutoffs who would reduce their hours of work to make themselves eligible for the new welfare program.

Table 8: Observed and Predicted Treatment Effects (TE) of the Vermont WRP on Various Outcomes

	Observed TE		Mechanical TE	Predicted TE MNL	MNL-FC	MixL	MixL-FC	MixL-Co	MixL-FC-Co
				Panel A: Labor Supply and Welfare Participation Outcomes					
All Sample									
Pr(p=1)	−3.3	(3.4)	8.7	3.5	3.2	2.0	2.1	−2.1	1.7
Pr(h>0)	−1.0	(3.7)	0.0	1.8	2.5	0.0	3.3	−0.7	3.8
E(h)	−1.1	(1.4)	0.0	0.1	0.6	0.0	0.9	−0.1	1.1
E(h|h>0)	−1.1	(1.2)	0.0	−0.6	−0.4	0.0	−0.3	0.2	−0.3
On Welfare									
Pr(h>0|p=1)	4.4	(4.7)	16.4	8.0	10.9	3.9	12.3	−6.3	13.6
E(h|p=1)	1.3	(1.0)	6.7	2.1	3.5	1.5	4.0	−0.8	4.5
E(h|p=1, h>0)	3.2	(2.9)	12.9	2.1	5.3	2.7	5.3	1.9	5.5
Off Welfare									
Pr(h>0|p=0)	−6.9	(3.9)	−3.4	0.0	0.0	−0.6	0.0	−0.3	−0.1
E(h|p=0)	−3.6	(1.7)	−1.3	0.0	0.0	0.0	0.0	−0.6	−0.1
E(h|p=0, h>0)	−1.3	(1.2)	−0.1	0.0	0.0	0.2	0.0	−0.6	0.0
				Panel B: Earnings and Income Outcomes					
All Sample									
E(Earnings)	5.7	(12.4)	0.0	0.1	3.6	0.0	5.7	−0.7	6.9
E(Income)	−1.9	(10.4)	7.3	6.8	6.8	3.1	7.1	−3.4	7.4
On Welfare									
E(Earnings|p=1)	5.5	(4.9)	35.5	14.5	21.7	9.5	24.8	−4.4	27.9
E(Income|p=1)	14.2	(6.1)	27.0	17.4	17.5	11.3	20.0	−0.1	22.3
Off Welfare									
E(Earnings|p=0)	−4.1	(15.7)	2.2	0.0	0.0	−0.1	0.0	−3.8	−0.4
E(Income|p=0)	−11.4	(15.6)	−3.9	0.0	0.0	−1.0	0.0	−4.2	−0.4
				Panel C: Cost-Benefit Outcomes					
All Sample									
E(Welfare benefits)	−4.1	(7.1)	7.3	5.2	1.3	3.0	−0.9	−2.4	−2.2
E(Tax payments)	0.6	(0.9)	0.0	−0.2	−0.1	0.0	−0.1	0.0	−0.1
E(EITC)	−2.9	(1.9)	0.0	1.2	1.8	0.1	2.3	−0.3	2.6
E(Net costs)	−7.6	(6.6)	7.4	6.6	3.2	3.1	1.4	−2.7	0.5
On Welfare									
E(Welfare benefits|p=1)	6.9	(5.8)	−20.7	−2.1	−11.4	−1.2	−13.0	6.2	−14.8
E(Tax payments|p=1)	0.0		0.0	0.0	0.0	0.0	0.0	0.0	0.0
E(EITC|p=1)	1.8	(1.8)	12.2	5.0	7.2	3.0	8.2	−1.9	9.2
E(Net costs|p=1)	8.7	(5.1)	−8.5	2.9	−4.2	1.8	−4.8	4.3	−5.6
Off Welfare									
E(Tax payments|p=0)	0.6	(1.3)	1.0	0.0	0.0	0.2	0.0	−0.1	0.0
E(EITC|p=0)	−6.8	(2.3)	−5.0	0.0	0.0	−0.8	0.0	−0.6	−0.1
E(Net costs|p=0)	−7.3	(2.9)	−6.0	0.0	0.0	−1.0	0.0	−0.4	0.0

Notes: Experimental treatment effect estimates and cross-state evaluation results. Robust standard errors of experimental estimates are in parentheses. Predictions based on the parameter estimates reported in Table 3. Net costs = Welfare benefits - Taxes + EITCs. Probability differences in percentage points. Weekly hours of work. Weekly costs in 1999 dollars. The mechanically calculated treatment effects are calculated under the assumptions: 1) the AFDC recipients would keep receiving welfare benefits under the new MFIP rules without changing their labor supply behavior; 2) the eligible non-participants under the AFDC rules would maintain their labor supply behavior and no take-up decisions; 3) the "windfall beneficiaries," who are not eligible for the AFDC but newly become eligible for the MFIP due to the MFIP's higher income cutoffs, would participate in the MFIP while holding their choice of working hours fixed; and 4) there would be no "switchers," individuals earning more than the MFIP income cutoffs who would reduce their hours of work to make themselves eligible for the new welfare program.

Appendix Table 1: Estimates of Wage Equation

	MN MFIP	MN AFDC	VT WRP	VT AFDC
Age 20-24	0.183	0.0707	0.255	0.0602
	(0.129)	(0.0618)	(0.150)	(0.104)
Age 25-34	0.102	0.109	0.354*	0.113
	(0.125)	(0.0597)	(0.160)	(0.0853)
Age 35-44	0.0678	0.103	0.266	0.0668
	(0.131)	(0.0754)	(0.160)	(0.0963)
Age 45+	−0.465	−0.158	0.445*	0.0426
	(0.505)	(0.227)	(0.199)	(0.158)
GED	0.160*	0.109	0.130	−0.0500
	(0.0620)	(0.0650)	(0.126)	(0.106)
HS diploma	0.206**	0.185**	0.0988	0.0177
	(0.0590)	(0.0569)	(0.0790)	(0.0694)
Tech/AA/2 yr college	0.192*	0.280**	0.0125	0.0500
	(0.0835)	(0.0716)	(0.132)	(0.130)
4 yr college	0.485**	0.367*	0.311*	0.107
	(0.122)	(0.142)	(0.127)	(0.150)
County-level	0.0433	−0.0739	−0.00367	−0.0191
unemployment rate	(0.0632)	(0.0427)	(0.0303)	(0.0365)
Constant	1.782**	2.127**	1.587**	1.845**
	(0.185)	(0.122)	(0.176)	(0.144)
Number of Obs	325	483	214	219

Notes: Robust standard errors in parentheses. * $p<0.05$, ** $p<0.01$

Appendix Table 2: Results of Estimating Models of Labor Supply and Welfare Participation, No Taste Shifters

	(1) MNL	(2) MNL-FC	(3) MixL	(4) MixL-FC	(5) MixL-Corr	(6) MixL-FC-Corr
Consumption/100	1.028**	0.761**	1.861**	0.778**	0.780	0.779**
	(0.128)	(0.120)	(0.936)	(0.128)	(0.610)	(0.133)
$(\text{Consumption}/100)^2$	0.035*	0.006	−0.247	0.006	−0.094	0.006
	(0.019)	(0.011)	(0.191)	(0.012)	(0.132)	(0.012)
Hours of work/10	−0.675**	1.484**	12.645	1.467**	10.667*	1.466**
	(0.128)	(0.244)	(7.892)	(0.246)	(6.089)	(0.246)
$(\text{Hours of work}/10)^2$	0.083**	−0.278**	−4.289	−0.276**	−3.777*	−0.276**
	(0.020)	(0.034)	(2.683)	(0.035)	(2.043)	(0.035)
(Consumption/100)	−0.188**	−0.088**	−0.018	−0.091**	0.267	−0.091**
×(Hours of work/10)	(0.038)	(0.027)	(0.238)	(0.028)	(0.284)	(0.029)
Welfare participation	−1.234**	−0.728**	−1.626*	−0.882**	−1.562**	−0.894**
	(0.189)	(0.227)	(0.922)	(0.322)	(0.686)	(0.321)
Fixed costs of work/100		6.791**		6.609**		6.600**
		(1.298)		(1.295)		(1.305)
Unobserved heterogeneity						
σ_h			26.394	0.010	20.863*	0.001
			(16.770)	(0.139)	(11.775)	(0.146)
σ_p			1.324	0.903	1.483*	0.938
			(1.197)	(0.718)	(0.898)	(0.756)
σ_{hp}					−0.865**	−0.554
					(0.232)	(1.299)
Log Likelihood	−1676.218	−1447.535	−1589.171	−1447.116	−1588.353	−1447.065
Number of Obs	753	753	753	753	753	753

Notes: Estimation based on data from the Minnesota control group. Number of individuals*Number of available choice alternatives = 8,538. Each individual faces up to 18 choice alternatives. Standard errors in parenthesis. * $p < 0.10$, ** $p < 0.05$.

Appendix Table 3: Actual and Predicted Probabilities for the Minnesota AFDC Sample

		Predictions					
	Actual	MNL	MNL-FC	MixL	MixL-FC	MixL-Co	MixL-FC-Co
h=0, p=0	13.01	6.79	12.51	10.92	12.61	13.73	12.68
h=0, p=1	27.09	17.63	27.56	23.38	27.31	22.93	27.12
h=10, p=0	2.12	11.61	1.37	8.64	1.30	9.48	1.30
h=10, p=1	1.99	10.66	1.07	5.31	1.18	3.78	1.17
h=20, p=0	5.71	15.05	7.26	12.29	7.08	9.75	7.08
h=20, p=1	0.66	1.76	1.11	1.26	1.25	1.73	1.30
h=30, p=0	9.30	12.84	16.55	12.00	16.51	11.16	16.52
h=30, p=1	0.40	0.14	0.30	0.18	0.31	0.19	0.33
h=40, p=0	32.27	9.21	18.23	9.83	18.40	10.27	18.41
h=40, p=1	0.13	0.05	0.17	0.07	0.16	0.03	0.18
h=50, p=0	4.52	6.23	10.30	7.02	10.36	7.42	10.37
h=50, p=1	0.00	0.02	0.06	0.03	0.06	0.01	0.07
h=60, p=0	1.73	3.99	3.01	4.49	2.98	4.65	2.98
h=60, p=1	0.00	0.01	0.01	0.01	0.01	0.00	0.02
h=70, p=0	0.27	2.46	0.45	2.56	0.44	2.65	0.44
h=70, p=1	0.00	0.01	0.00	0.00	0.00	0.00	0.00
h=80, p=0	0.80	1.54	0.04	2.00	0.03	2.22	0.03
h=80, p=1	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Notes: Actual sample fractions and mean predicted probabilities. Predictions based on the parameter estimates reported in Table 3. Number of observations=753.

Appendix Table 4: Counterfactual Outcomes of the Minnesota AFDC Sample with MFIP Benefits

	Actual		Mechanical Calculations	Predictions											
				MNL		MNL-FC		MixL		MixL-FC		MixL-Co		MixL-FC-Co	
	AFDC	MFIP	MFIP	AFDC	MFIP	AFDC	MFIP	AFDC	MFIP	AFDC	MFIP	AFDC	MFIP	AFDC	MFIP
Panel A: Labor Supply and Welfare Participation Outcomes															
All Sample															
Pr(p=1)	30.28	52.88	54.58	30.28	42.79	30.28	39.25	30.24	35.77	30.29	36.14	28.67	25.17	30.19	34.94
Pr(h>0)	59.89	57.01	59.89	75.58	80.41	59.93	65.98	65.70	65.70	60.08	68.50	63.34	61.45	60.19	69.98
E(h)	21.96	19.62	21.96	21.67	21.56	21.67	23.16	21.34	21.36	21.72	24.15	21.28	21.09	21.76	24.73
E(h|h>0)	36.67	34.52	36.67	28.41	26.68	36.01	35.01	32.11	32.14	35.99	35.17	33.37	34.06	35.99	35.25
On Welfare															
Pr(h>0|p=1)	10.53	40.82	50.36	40.31	67.82	7.56	38.68	21.68	34.48	8.57	42.65	17.35	7.33	8.99	46.30
E(h|p=1)	1.61	11.45	17.33	4.59	11.37	1.33	10.91	2.66	7.00	1.52	12.15	2.25	1.67	1.64	13.51
E(h|p=1, h>0)	15.33	28.04	34.42	10.69	16.80	12.24	26.77	11.00	19.86	12.27	27.00	11.28	22.12	12.35	27.62
Off Welfare															
Pr(h>0|p=0)	81.33	75.19	71.35	89.64	89.64	81.59	81.59	83.93	82.43	81.52	81.52	80.80	79.5	81.47	81.20
E(h|p=0)	30.80	28.79	27.53	28.67	28.67	30.15	30.15	29.03	28.84	30.19	30.19	28.64	27.41	30.17	30.02
E(h|p=0, h>0)	37.87	38.48	38.59	31.97	31.97	36.94	36.94	34.50	34.87	37.00	37.00	35.37	34.41	37.00	36.93
Panel B: Earnings and Income															
All Sample															
E(Earnings)	212.14	183.88	212.14	208.31	204.98	207.70	218.60	199.63	199.74	206.54	225.73	196.48	195.02	206.24	229.95
E(Income)	279.01	284.35	295.74	276.99	296.88	277.22	295.72	267.83	274.63	276.10	297.55	259.82	252.91	275.63	298.79
On Welfare															
E(Earnings|p=1)	8.45	85.81	134.11	38.98	101.48	8.71	90.33	20.85	58.89	9.98	100.64	17.97	15.48	10.57	111.56
E(Income|p=1)	199.95	269.08	293.21	214.09	283.92	195.03	258.44	202.64	239.36	195.79	265.93	200.43	201.29	196.15	273.17
Off Welfare															
E(Earnings|p=0)	300.61	293.93	305.92	266.21	266.21	281.19	281.19	266.92	266.34	281.44	281.44	261.43	251.35	281.29	279.99
E(Income|p=0)	313.35	301.47	298.78	289.05	289.05	300.84	300.84	285.66	282.63	300.99	300.99	276.67	266.06	300.83	299.53
Panel C: Cost-Benefit Outcomes															
All Sample															
E(Welfare benefits)	57.01	86.04	73.74	50.17	68.44	56.55	59.53	54.07	60.60	56.24	53.09	51.25	46.67	55.89	49.35
E(Tax payments)	10.53	7.86	10.53	9.77	8.55	9.69	9.22	9.81	9.77	9.51	9.31	10.06	10.08	9.45	9.33
E(EITC)	20.39	22.29	20.39	28.28	32.02	22.66	26.80	23.95	24.07	22.84	28.04	22.15	21.30	22.96	28.83
E(Net costs)	66.87	100.47	83.59	68.68	91.90	69.52	77.11	68.20	74.90	69.56	71.82	63.34	57.89	69.40	68.84
On Welfare															
E(Welfare benefits|p=1)	188.30	162.71	135.10	160.49	152.28	183.06	148.52	174.01	165.75	182.08	143.69	175.64	182.80	181.63	138.30
E(Tax payments|p=1)	0.00	0.61	1.55	0.00	0.33	0.00	0.76	0.00	0.42	0.00	0.86	0.00	0.16	0.00	1.01
E(EITC|p=1)	3.21	21.18	25.55	14.62	30.49	3.25	20.35	7.79	15.13	3.73	22.46	6.81	3.18	3.95	24.32
E(Net costs|p=1)	191.50	183.28	159.09	175.11	182.44	186.31	168.10	181.80	180.46	185.82	165.29	182.45	185.81	185.58	161.61
Off Welfare															
E(Tax payments|p=0)	15.11	15.99	21.33	12.12	12.12	12.34	12.34	12.83	13.53	12.36	12.36	13.24	12.92	12.35	12.28
E(EITC|p=0)	27.85	23.53	14.19	34.96	34.96	31.99	31.99	31.57	29.83	31.90	31.90	28.48	27.63	31.88	31.82
E(Net costs|p=0)	12.74	7.54	−7.14	22.84	22.84	19.66	19.66	18.74	16.30	19.54	19.54	15.24	14.71	19.53	19.54

Notes: Actual outcomes and counterfactual policy simulation results. Predictions based on the parameter estimates reported in Table 3. The mechanical calculations are done the assumptions: 1) the AFDC recipients would keep receiving welfare benefits under the new MFIP rules without changing their labor supply behavior; 2) the eligible non-participants under the AFDC rules would maintain their labor supply behavior and no take-up decisions; 3) the "windfall beneficiaries," who are not eligible for the AFDC but newly become eligible for the MFIP due to the MFIP's higher income cutoffs, would participate in the MFIP while holding their choice of working hours fixed; and 4) there would be no "switchers," individuals earning more than the MFIP income cutoffs who would reduce their hours of work to make themselves eligible for the new welfare program.

Appendix Table 5: Counterfactual Outcomes of the Vermont AFDC Sample with WRP Benefits

			Mechanical	Predictions											
	Actual		Calculations	MNL		MNL-FC		MixL		MixL-FC		MixL-Corr		MixL-FC-Corr	
	AFDC	WRP	WRP	AFDC	WRP	AFDC	WRP	AFDC	WRP	AFDC	WRP	AFDC	WRP	AFDC	WRP
Panel A: Labor Supply and Welfare Participation Outcomes															
All Sample															
Pr(p=1)	36.41	33.07	45.11	39.10	42.57	36.05	39.24	33.45	35.41	33.38	35.47	26.70	24.58	32.43	34.12
Pr(h>0)	55.43	54.40	55.43	70.73	72.52	56.66	59.13	65.73	65.74	59.10	62.35	65.28	64.56	60.01	63.76
E(h)	19.36	18.28	19.36	20.70	20.78	20.32	20.94	21.79	21.80	21.21	22.15	21.60	21.49	21.56	22.70
E(h|h>0)	34.92	33.77	34.92	29.03	28.48	35.68	35.25	32.72	32.72	35.69	35.37	32.85	33.03	35.73	35.44
On Welfare															
Pr(h>0|p=1)	14.93	19.35	31.33	44.68	52.71	15.25	26.19	30.19	34.11	17.70	29.99	20.50	14.20	19.05	32.69
E(h|p=1)	2.45	3.79	9.17	6.10	8.21	3.26	6.75	4.61	6.06	3.82	7.80	3.38	2.62	4.20	8.69
E(h|p=1, h>0)	16.40	19.58	29.27	13.45	15.55	18.32	23.57	14.53	17.18	18.42	23.74	15.17	17.08	18.67	24.18
Off Welfare															
Pr(h>0|p=0)	78.63	71.71	75.25	86.94	86.94	79.01	79.01	83.02	82.46	79.04	79.04	80.66	80.34	78.99	78.88
E(h|p=0)	29.04	25.44	27.73	29.75	29.75	29.57	29.57	29.96	29.92	29.60	29.6	27.95	27.35	29.61	29.54
E(h|p=0, h>0)	36.93	35.67	36.85	34.21	34.21	37.39	37.39	35.95	36.13	37.40	37.40	34.51	33.92	37.43	37.40
Panel B: Earnings and Income															
All Sample															
E(Earnings)	138.20	143.88	138.20	142.11	142.26	141.22	144.78	149.74	149.76	146.28	151.99	148.06	147.34	148.03	154.94
E(Income)	229.00	227.08	236.35	235.00	241.75	233.48	240.26	233.85	236.92	233.89	241.03	220.33	216.89	233.84	241.28
On Welfare															
E(Earnings|p=1)	12.52	17.98	48.00	39.21	53.72	18.17	39.9	27.81	37.29	21.27	46.11	20.44	16.02	23.23	51.16
E(Income|p=1)	209.31	223.49	236.30	225.29	242.67	211.42	228.93	217.87	229.19	213.21	233.23	213.09	212.98	214.30	236.6
Off Welfare															
E(Earnings|p=0)	210.17	206.09	212.33	200.20	200.2	203.03	203.03	204.24	204.18	203.33	203.33	190.62	186.86	203.24	202.86
E(Income|p=0)	240.28	228.85	236.40	231.57	231.57	237.42	237.42	233.89	232.87	237.74	237.74	218.36	214.15	237.63	237.2
Panel C: Cost-Benefit Outcomes															
All Sample															
E(Welfare benefits)	69.89	65.75	77.24	68.85	74.03	68.49	69.8	61.58	64.57	62.54	61.63	50.10	47.68	60.21	58.05
E(Tax payments)	4.12	4.70	4.12	3.94	3.74	3.41	3.33	4.59	4.59	3.45	3.40	4.58	4.58	3.45	3.40
E(EITC)	25.03	22.15	25.03	27.98	29.20	27.19	29.01	27.12	27.18	28.51	30.81	26.75	26.45	29.05	31.69
E(Net costs)	90.80	83.19	98.15	92.89	99.49	92.27	95.48	84.11	87.16	87.61	89.04	72.27	69.55	85.81	86.34
On Welfare															
E(Welfare benefits|p=1)	191.93	198.84	171.24	171.45	169.34	186.49	175.06	179.70	178.52	184.00	170.98	184.91	191.10	182.42	167.60
E(Tax payments|p=1)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
E(EITC|p=1)	4.85	6.68	17.06	14.63	19.60	6.76	13.97	10.36	13.38	7.93	16.14	7.74	5.85	8.65	17.83
E(Net costs|p=1)	196.78	205.52	188.29	186.08	188.95	193.26	189.03	190.06	191.90	191.94	187.12	192.65	196.96	191.07	185.43
Off Welfare															
E(Tax payments|p=0)	6.48	7.03	7.51	5.68	5.68	4.67	4.67	6.25	6.41	4.68	4.68	5.89	5.79	4.68	4.67
E(EITC|p=0)	36.59	29.79	31.58	37.06	37.06	39.06	39.06	35.90	35.10	39.09	39.09	33.63	33.08	39.07	39.01
E(Net costs|p=0)	30.10	22.76	24.07	31.37	31.37	34.39	34.39	29.65	28.69	34.41	34.41	27.74	27.29	34.39	34.35

Notes: Actual outcomes and counterfactual policy simulation results. Predictions based on the parameter estimates reported in Table 3. The mechanical calculations are done the assumptions: 1) the AFDC recipients would keep receiving welfare benefits under the new MFIP rules without changing their labor supply behavior; 2) the eligible non-participants under the AFDC rules would maintain their labor supply behavior and no take-up decisions; 3) the "windfall beneficiaries," who are not eligible for the AFDC but newly become eligible for the MFIP due to the MFIP's higher income cutoffs, would participate in the MFIP while holding their choice of working hours fixed; and 4) there would be no "switchers," individuals earning more than the MFIP income cutoffs who would reduce their hours of work to make themselves eligible for the new welfare program.