Validation of the Discrete Choice Labor Supply Model by Methods of the New Tax Responsiveness Literature*

by

Thor O. Thoresen§ and Trine E. Vattø†

Abstract

The static structural discrete choice labor supply model continues to be a workhorse in the process of policy-making, extensively used by policy-makers to predict labor supply effects of changes in the personal income tax system. A widely used alternative to obtain estimates of individual tax responsiveness is to exploit the diversity of tax treatment generated by a tax reform to recover tax induced outcome differences in data. Response estimates obtained from analysis of tax reforms are less useful for describing effects of prospective policies, but represent an underexploited source of information for out-of-sample validation of labor supply models. The present study describes how estimates of responses in working hours and income, generated from a tax reform, can be used to validate a discrete choice labor supply model; thus, bringing together and providing guidance to how results of two main avenues of obtaining estimates of tax responsiveness can be compared and interpreted. We find that the discrete choice model used by Norwegian policy-makers performs well as measured by this type of validation.

Keywords: Model validation, Tax response, Discrete choice labor supply model, Elasticity of earned income
JEL classification: H21, H24, H31, J22

*Financial support from the Research Council of Norway is gratefully acknowledged. We have received valuable comments on earlier versions of the paper from Karl Ove Aarbu, John Dagsvik, Frank Fossen, Zhiyang Jia, Jukka Pirtillä, Arvid Raknerud, Victoria Sparrman, Viktor Steiner and seminar participants at Wirtschaftspolitisches Seminar in Berlin, Skatteforum (Moss, June, 2011), the 67th IIPF-conference (Ann Arbor, August, 2012), Workshop on Recent Advances in Labor Supply Modeling (Dublin, May, 2012), Workshop in Public Economics (Uppsala, June, 2012), and the Melbourne Institute Seminar Series (December, 2013). We would also like to thank the Editor and two anonymous referees for their helpful comments and stimulating criticism of an earlier draft.

§ Statistics Norway, Oslo Fiscal Studies (Department of Economics, University of Oslo), and CESifo (Munich), email: tot@ssb.no
† Corresponding author. Statistics Norway, email: tev@ssb.no

1. Introduction

Some institutions, such as the Joint Committee on Taxation (U.S.), the Institute for Fiscal Studies (U.K.), and the Research Department of Statistics Norway, are expected to deliver empirical estimates of effects to the decision-makers in their respective countries. The application of certain modeling tools is often a prerequisite for this, and the structural static labor supply model represents a practical alternative for predicting effects of tax changes on the labor market behavior of income earners. Based on cross-sectional observations of households' and individuals' consumption and connections to the labor market (typically working hours), labor supply models can be estimated and then used in the policy-making process for simulations of short term labor market effects of prospective changes in the tax system.

In the category of structural labor supply modeling approaches, the discrete choice model of labor supply based on the random utility modeling approach (van Soest, 1995) stands out, as it has gained widespread popularity among public finance practitioners (Creedy and Kalb, 2005). For example, Norwegian decision-makers have access to a discrete choice labor supply model through the model system LOTTE (Aasness, Dagsvik and Thoresen, 2007).

However, concerns have been raised about the ability of structural models to generate robust predictions about the effects of policy changes, see for example LaLonde (1986) and Imbens (2010). As models may be too stylized or may suffer from misspecification, predictions of effects of counterfactual policy alternatives are not always trustworthy. The use of predictions from structural models as input to the policy-making process is therefore disputed, and the policy analyst may resort to providing alternative and less detailed information about tax responsiveness, such as tax response estimates obtained from studies using quasi-experimental econometric designs. In the present study we argue that instead of dismissing the structural labor supply model approach as a tool for policy-making completely, more effort should be put into qualifying models through validation. In this perspective, results from experiments serve as useful information sources for validation of prediction models (Blundell, 2006; Keane, 2010a).

Models should be assessed with respect to realism and reasonability of assumptions. As for model performance, the researcher usually does not have much information apart from goodness-of-fit measures. Such evidence is valuable, but insufficient, and a key test of model validity is to examine how well the model predicts out-of-sample labor supply behavior. Some of the most prominent examples of the power of structural discrete choice modeling, such as the McFadden's predictions of the use of a new train transportation system in the Bay Area of San Francisco (McFadden, Talvitie et al., 1977) and the model developed to evaluate the Mexican PROGRESA school subsidy program (Todd and Wolpin, 2006), have received their status because of extensive out-of-sample model validations.

There are several alternatives for out-of-sample validations of the discrete choice model. Mechanical use of experimental sources for validation is problematic, as they are informative about the combined impact of the policy change in question and other effects, such as contemporaneous changes in the tax and benefit and welfare systems and the business cycle. In this perspective, the so-called elasticity of taxable income approach (the ETI approach), or interchangeably labelled the new tax responsiveness literature (the NTR approach), represents a promising alternative for use in external validations, as it denotes a well-established procedure to rinse out the effects of taxes. Studies of the large and growing ETI/NTR literature exploit that tax reforms generate net-of-tax rate changes along the income scale, often resulting in substantial tax changes for some tax-payers, whereas others are more or less unaffected. Taxable income is used as the main measure of outcome in this literature, as it in principle captures all the public policy relevant behavioral responses of a reform (hours worked, effort, tax avoidance and evasion, change of job, etc). The review of this literature in Saez, Slemrod, and Giertz (2012) clearly reveals that this has been a fertile field of research in recent decades, even though there are well-known methodological weaknesses involved.

Here we suggest using the ETI/NTR approach to validate the discrete choice labor supply model. However, in the validation we shall use estimates of responses in working hours and earned income, and not responses in total taxable income. As for the terminology, the use of the acronym ETI for "elasticity of taxable income approach" may therefore be less suitable in the present context. To maintain that we use exactly the same techniques as studies under the ETI label, but to avoid the potential distraction that comes from the reference to "taxable income", we will employ the other term which we see used for the labelling of this type of studies: the NTR approach, an expression introduced by Goolsbee (1999).[1]

Thus, results of probably the two most important sources of information on tax responsiveness are brought together in the present study: simulation results from the discrete choice labor supply model, estimated on a single cross-section of data, and estimates obtained from analysis of panel data and when tax reforms are used for identification, hereafter referred to as the NTR approach. Estimates from "natural experiments" have limited value in a prediction context (less external validity), because they rely on a particular reform for identification, and parameters are therefore not usually policy invariant, but the NTR approach represents a powerful and underexploited tool in a validation context. Of course, this exercise cannot prove the model "correct", but is helpful in detecting misspecified models.

The main contribution of the present study is to show how results of the two techniques can be understood and utilized in a validation context. We use a very large dataset for Norwegian wage earners, based on administrative registers, and exploit the tax changes due to the Norwegian tax

[1] Thus, we use the NTR acronym only for "semantical" reasons, as the ETI approach and the NTR approach refer to exactly the same literature: studies that report responses in taxable income with respect to changes in marginal tax rates.

reform of 2006 to obtain two sets of tax response estimates for wage earners (separately for single females, single males, and females and males in couples): one set of NTR elasticities for working hours and one set for earned income. Then the discrete choice labor supply model is estimated on the same data, and results from model simulations of the 2006-reform are recalculated into NTR elasticities for working hours. The description of the conversion of results from the random utility discrete choice model into NTR results is a key contribution of the paper.

Another main contribution of the paper comes from having access to panel data information for both working hours and earned income, which means that we are able to elaborate upon key characteristics of the discrete choice model in a validation perspective. The conventional discrete choice model (van Soest, 1995) implies that the individual specific wage is kept fixed in the transition from pre-reform to post-reform tax schedules. In contrast, in the standard NTR approach, which focuses on responses in income, one may also see responses in wages (in addition to changes in hours of work), as individuals may react to a tax change by finding a new job, take on other tasks in the present job, or change behavior in the wage bargaining, etc., see Feldstein (1995). Thus, if we observe substantially larger NTR responses in earned income than in working hours, that may call for other modeling tools. One could think of allowing for specific relationships between working hours and wage rates in the model simulations,but we are not aware of any simulation model, with a similar design as ours, that includes a specific tied relationship between working hours and wage rates in the simulations. An alternative is to let the wage be determined by a suitable “after-model” to account for general equilibrium effects on wages, see Creedy and Duncan (2005) and Peichl and Siegloch (2012).[2]

Moreover, and related to the question of different margins of tax response, a discussion of the relationship between responses in earnings and working hours is also useful for future validation practice, in that it provides guidance on the use of income information alone in a validation exercise like the present one. Large register-based datasets on income are now commonly accessible for the analyst, in the Nordic countries and in several other countries (UN, 2007). Nonetheless, income information is usually more accessible than data on hours of work.

The paper is organized as follows. In Section 2 we present the two methodological approaches for obtaining tax response estimates, whereas Section 3 presents some studies of the validation literature. Section 4 describes the data sources we have utilized in this study, gives a brief overview over the tax reform of 2006, and shows preliminary data descriptions, given the main characteristics of the reform and their expected implications for income patterns. In Section 5 we present the results of the validation exercise, and Section 6 concludes the paper.

[2] There are studies accounting for interrelationships between wages and preferences in the estimation of the model, see for example Moffitt (1984) and Blundell and Shephard (2012). Dagsvik and Jia (2014) discuss identification issues in a setting when there is unobserved heterogeneity in the wage equation and where tax-payers have preferences for jobs (which is a reasonable extension of the standard discrete choice model if one would like to accommodate for effects through wages).

2. Two approaches to obtain estimates of short term tax responsiveness

A whole range of different tax response estimates can be found in the labor supply literature, reflecting inter alia different theoretical models and methodological approaches. In the present analysis, we discuss evidence from two well-known static approaches to produce short term measures of tax responsiveness: tax simulation based on a structural discrete choice labor supply model, and reduced-form estimation exploiting differential changes in tax treatment following from tax reforms. Given that estimation of structural labor supply models often involves severe econometric challenges,[3] see reviews in Blundell and MaCurdy (1999), Kniesner and Ziliak (2008), Meghir and Phillips (2010) and Keane (2011), the NTR approach, involving standard panel data techniques, may represent a more convenient empirical approach for the practitioner of public finance. However, NTR estimates are not usually invariant to the policy change that have been used to estimate them (Blundell and MaCurdy, 1999; Heckman and Vytlacil, 2005; Chetty, 2009), and cannot replace a well-behaving structural model in a (general) prediction context. But the NTR approach provides important and underexploited information for out-of-sample validation of structural labor supply models, as also argued by Blundell (2006) and Keane (2010a).

In this section, we present the main characteristics of the two methods of deriving individual tax response estimates. First, a discrete choice labor supply model is presented, and then we describe how tax response estimates can be derived when making use of individual panel data over a reform period.

2.1 The discrete choice labor supply model

Discrete choice models of labor supply, based on the random utility modeling approach, have gained widespread popularity,[4] mainly because they are much more practical than the conventional continuous approach based on marginal calculus; see Creedy and Kalb (2005) for a survey of the literature and van Soest (1995), Duncan and Giles (1996), Bingley and Walker (1997), Blundell et al. (2000), van Soest, Das and Gong (2002), Creedy, Kalb and Scutella (2006), Haan and Steiner (2005), Labeaga, Olivier and Spadaro (2008), and Blundell and Shephard (2012) for applications. With the

[3] It can be argued that the discrete choice version of structural modeling is a more practical method than the conventional continuous approach, based on marginal calculus. The structural labor supply model associated with Hausman (Hausman, 1985) becomes very complicated when more general and flexible model specifications are used, see Bloemen and Kapteyn (2008).

[4] Despite its popularity among practitioners of labor supply analysis, less attention is devoted to this framework in recent reviews of the literature. Keane (2011), for example, essentially ignores the (static) discrete choice approach to labor supply altogether.

discrete choice approach, it is easy to deal with nonlinear and nonconvex economic budget constraints, and to apply rather general functional forms of the utility representation.

With particular distributional assumptions about the stochastic disturbances in the utility function one can derive tractable expressions for the distribution of hours of work, such as the multinomial logit model or the nested multinomial logit model. The maximization problem for a person in a single-individual household can be seen as choosing between bundles of consumption (*C*) and leisure (*L*), subject to a budget constraint, $C = f\left(hw, I\right)$, where *h* is hours of work, *w* is the wage rate, *I* is non-labor income, *C* is (real) disposable income and $f(\cdot)$ is the function that transforms gross income into after-tax household income.

The utility function of the household is assumed to be additively separable, $U\left(C,h\right) = v\left(C,h\right) + \varepsilon(C,h)$, where $v\left(\cdot\right)$ is a positive deterministic function and ε the random unobserved components for individual *i* and choice *j*. We assume that the random components are i.i.d. extreme value distributed with c.d.f. $\exp\left(\exp\left(-x\right)\right)$ for positive *x*, which implies independence of irrelevant alternatives (IIA). The strict IIA assumption can be weakened, however, by allowing for random effects in utility parameters or in relation to the wage rate.[5]

Let $v\left(\cdot\right)$ be the representative utility of jobs with hours of work *h*, a given individual specific wage rate *w*, and non-labor income *I*. By applying standard results in discrete choice theory (McFadden, 1984), it follows that the probability that the agent will choose working hours *h* can be expressed as

$$(2.1) \qquad P\left(h\right) = \frac{\exp v(f\left(hw,I\right),h)}{\exp v(f\left(hw,I\right),0) + \sum_{h\in D} \exp v(f\left(hw,I\right),h)}.$$

We see different specifications of the deterministic part in the literature.[6] Here, we use a flexible Box-Cox functional form specification, $v(C,h) = \alpha_0 \frac{\left(C-C_0\right)^{\alpha_1} - 1}{\alpha_1} + (\beta_0 + \gamma Z)\frac{\left(\bar{h}-h\right)^{\beta_1} - 1}{\beta_1}$, where *C* measures the household-adjusted consumption level, constructed by dividing the couple or individual's disposable income by $\sqrt{N}$, where *N* is the number of individuals in the household (including children under 18). An additional interaction term between consumption and leisure has negligible effect and is dropped. C_0 represents the minimum or subsistence household-adjusted

[5] We replace the wage rate by a wage equation that includes a stochastic error term, and thus a mixed multinomial logit model follows, see McFadden and Train (2000) and Haan (2006).

[6] Quadratic or translog functional forms for the systematic part of the utility function have also been used in several applications, see, e.g., van Soest (1995). One advantage of the Box-Cox functional form is that it is globally monotone in consumption and leisure; see Dagsvik et al. (2014) for a discussion of this issue. In practice, the choice of functional form seems to have little impact on results.

consumption level, here set to 60,000 Norwegian kroner (NOK), or approximately 8,900 US dollars and 7,200 Euros.[7] $\bar{h}$ is defined as 80 hours per week and *h* is working hours per week, so that $\left(\bar{h}-h\right)$ measures leisure time. *X* is a vector of taste-modifying variables, including age and number of children.

To improve the fit to data, researchers often have resorted to specifications where the systematic term of the utility function has been modified by introducing alternative specific constant terms; see van Soest (1995). This can for instance be rationalized by a model set-up where individuals have preferences over jobs while allowing for certain restrictions in the choice set (Aaberge, Dagsvik and Strøm, 1995; Aaberge, Colombino and Strøm, 1999; Dagsvik and Strøm, 2006; Dagsvik and Jia, 2014; Dagsvik et al., 2014), in which the representative utility terms, $v(\cdot)$, or rather $\exp v(\cdot)$, are weighted by the frequencies of available jobs, $m(h)$. See Appendix A for a more detailed exposition of this model. At this stage, note that the empirical specification of this (latent) job choice model is similar to the model of van Soest (1995); it provides a rationalization for the dummy variables which in practice usually are added to the systematic part of the utility function.

Note that for any reasonable functional form one can obtain a perfect fit to cross-sectional data by choosing a sufficiently flexible specification, see also Train (2009) and Haan (2006). To achieve identification it is usually assumed that, after controlling for some individual characteristics, the parameters are constant across the population. There are additional complications with respect to identification following from the job choice specification, which are further discussed in Dagsvik and Jia (2014) and Dagsvik et al. (2014).

This modeling approach is utilized in the labor supply module of the Norwegian micro simulation model system LOTTE (Aasness, Dagsvik and Thoresen, 2007), a collection of models which are extensively used by policy-makers. The present version of the labor supply module (the so-called LOTTE-Arbeid model) is estimated on data from the Labour Force Survey; a general documentation of this data source can be found in Statistics Norway (2003). However, the model we shall validate in the following is instead estimated on data from the Wage Statistics (Statistics Norway, 2006), which is a much larger panel data set based on administrative registers; see more detailed data description in Section 4. The main reason for using this data source in the present discussion is that the panel dimension of the Wage Statistics implies that we can use NTR panel data techniques on the same data set, which is obviously advantageous from a validation perspective.

Appendix A presents estimation results for single males, single females, and, separately, for males and females in couple (married/cohabiting). They are utilized in the simulation of labor supply responses to the Norwegian tax reform of 2006, presented in Section 5.

[7] We use exchange rates of one US dollar for 6.74 Norwegian kroner (NOK) and one Euro for 8.37 NOK; both refer to average exchange rates in 2004.

Given the ambition to use the NTR method in an out-of-sample validation of model simulation results, we sum up some of the main features of the discrete choice labor supply model, which may prove important in the comparison. Firstly, given our partial approach, there is no wage response when altering the tax schedule. As we here provide information on NTR approach responses in both working hours and income, we shall return to the "fixed wage" assumption.

Secondly, we note that the whole tax function enters into the budget constraint of our labor supply model, which makes it easy to deal with nonlinear and nonconvex budgets. As we will soon see, the NTR approach is based on more conventional marginal criteria.

Thirdly, in a validation exercise it is important to be aware of that the random utility model foundation differs from the reasoning behind identification in the NTR approach. In the simulation of responses to tax changes in the discrete choice model one accounts for both the deterministic part and the unobservables, with error terms drawn from the relevant distribution. Here, responses are calculated at the individual level, keeping random error terms constant before and after the policy change, and then recalculated into overall responses by taking averages.[8] This means that we assume that the random error terms that represent the effect of unobservables in preferences change very slowly, and a reasonable approximization is therefore to assume that they are constant over the period under consideration.

Fourthly, as the model is nonlinear, responses differ substantially along the income scale and show strong dependence on the actual policy change involved. Given that we use a particular tax change in the validation of the model (the 2006 tax reform), this feature of the model is highlighted by the present analysis.

2.2 Response identification according to the new tax responsiveness literature

The NTR literature looks at changes to taxable income or gross income rather than hours of work or earned income, to seize all the policy relevant behavioral adjustments. After initial contributions by Lindsey (1987) and Feldstein (1995), the NTR framework has been utilized to obtain tax responses from tax reforms in a number of countries;[9] Saez, Slemrod, and Giertz (2012) provide a survey of the literature. Most studies present (uncompensated) net-of-tax rate elasticities for taxable income or gross income, which reflect a range of intensive margin[10] responses to the tax reform under study.

[8] The simulation procedure of Creedy and Kalb (2005) is an alternative: based on a specific drawing procedure each individual's pre-reform and post-reform probability distributions are determined and forms the basis for calculating average measures, before and after the policy change. The procedure seen in Kornstad and Thoresen (2006) uses the sample information on probabilities, ignoring individual level information about error terms, which can be justified by assuming that error terms are unknown to the agents themselves. See also Duncan and Weeks (2000).

[9] Two influential studies using US data are Auten and Carroll (1999) and Gruber and Saez (2002). Aarbu and Thoresen (2001) is a previous study using Norwegian data (tax changes according to the 1992 tax reform).

[10] As income growth is the dependent variable, extensive margin effects are usually not considered in the NTR literature.

Therefore, almost by definition, there are fewer NTR studies using earned income and working hours as dependent variables. However, Singleton (2011), Kleven and Schultz (2014) and Gelber (2014) are examples of studies using the NTR technique on responses in earned income, whereas Moffitt and Wilhelm (2000) discuss responses in working hours in a NTR setting. There is a closely related literature using tax reforms and quasi-experimental identification techniques, see for example Eissa and Liebman (1996) and Eissa and Hoynes (2006).[11]

We have recently seen discussions in the literature concerning the advantages of structural modeling versus results derived from quasi-experimental research designs; see, for instance, Chetty (2009), Angrist and Pischke (2010), Deaton (2010), Heckman (2010), Heckman and Urzua (2010), Imbens (2010), and Keane (2010a; 2010b). As Chetty (2009) emphasizes, the NTR methodology is not easy to place in relation to the two stereotype classifications, since it shares important characteristics with both strands of the literature.[12] For instance, like structural models, the NTR framework departs from an underlying utility-maximizing behavior and produces precise statements about welfare implications. The identification strategy has, however, important similarities with experimental studies, as tax reforms are used for the identification of the parameter of interest.

The approach taken in much of the NTR literature departs from an underlying utility-maximizing behavior similar to that seen in the standard labor supply literature above (Feldstein, 1999; Blomquist and Selin, 2010; Saez, Slemrod and Giertz, 2012). Individuals are assumed to maximize a utility function that increases in consumption (*C*) and decreases in taxable income (*q*), subject to a budget constraint described by $C = (1-\tau)q + R$, where τ is the marginal tax rate (which applies to a linear segment of the tax schedule), and *R* is virtual income. Accordingly, the "supply function" of taxable income is estimated as a function of the marginal tax rate and virtual income. The formulation thus suggests a closer relationship to the part of the structural labor supply literature that is based on estimation of a continuous labor supply function with a piecewise-linear budget constraint, as in Burtless and Hausman (1978), and Hausman (1985).[13]

Panel data covering a period of net-of-tax rate variation across individuals and across time (often covering a tax reform) have been the main data source for the identification of responses in the empirical framework of the NTR approach. Taxable income for individual *i* at time *t*, q_{it}, is explained by a time-specific constant, κ_t, the net-of-tax rate, $\log(1-\tau_{it})$, unobserved heterogeneity μ_i and the remaining iid error term, ξ_{it},

[11] Not to mention the extensive literature which uses other "experiments" (not tax reforms) to identify policy effects on working hours, see, for example, the references in Angrist and Pischke (2009).

[12] Chetty therefore introduces a third class, the "sufficient statistic" category, which covers studies that make predictions about welfare without estimating or specifying structural models.

[13] The Hausman approach thus deviates from the standard discrete choice model (van Soest, 1995), in which estimation is carried out directly on the utility specification.

(2.2) $\log q_{it} = \kappa_t + \lambda \log(1-\tau_{it}) + \mu_i + \xi_{it}$.

The basic framework for identification in the NTR literature consists of various estimations of a first-differenced version of (2.2), using panel data for two periods,[14]

(2.3) $\Delta \log q_i = \kappa + \lambda \Delta \log(1-\tau_i) + \Delta \xi_i$.

The coefficient of interest, λ , measures the elasticity of income with respect to changes in the net-of-tax rate defined as $\frac{1-\tau}{q}\frac{\partial q}{\partial(1-\tau)}$. The reliability of results depends on carefully framed empirical designs for the identification of the key parameter, including controls for individual characteristics that might affect income growth. One obvious methodological identification challenge (w.r.t. λ) has been the endogeneity of the tax rate, which has led to the estimation of (2.3) using IV techniques. For instance, Feldstein (1995) employs the difference-in-differences estimator, and let the change in the net-of-tax rates and the allocation into groups (groups more or less treated by the US tax reform of 1986) be determined by pre-reform income levels. Many post-Feldstein studies employ a closely related exclusion restriction, namely the change in net-of-tax rates based on a fixed first period income as instrument in an IV regression; see Auten and Carroll (1999) and Gruber and Saez (2002). Thus, the NTR literature is related to methods commonly used in the "experimentalist" or "program evaluation" literature. However, the conventional identification technique of the NTR literature implies that one is far from an ideal randomized trial situation.

The estimated elasticity can be interpreted as the average treatment effect of the treated. In other words, if we let a parameter δ be a zero-one indication of being treated (experiencing net-of-tax rates changes, or not),[15] we identify $E(\lambda | \delta_{it} = 1)$. This parameter is subject to conventional sample selection biases and cannot in general be used to simulate policy responses (Blundell and MaCurdy, 1999).[16] However, as we shall see, the method is useful in order to deliver tax response estimates to be used in a validation of a structural model.

[14] Repeated cross-sections can also be used in the estimation of this model, for example by addressing information on groups of tax-payers, before and after a reform. See Holmlund and Söderström (2011) and Vattø (2014) for studies introducing dynamics in the model specification.

[15] The Norwegian tax reform of 2006 can be given a dichotomous representation.

[16] The estimated elasticities can only be used to simulate hypothetical tax reforms under the assumption that the elasticity is constant over the income distribution, which is clearly not consistent with findings from the structural labor supply literature.

3. Previous validation studies

Model evaluations should include assessments in terms of model realism and reasonability of assumption, in addition to goodness of fit tests; on the latter, see, for example Train (2009). Some of the best examples of the strength of the structural discrete choice tool for economic planning, such as McFadden's predictions of effects of BART (McFadden, Talvitie et al., 1977) and the prediction model developed by Todd and Wolpin (2006) through the PROGRESA project, have obtained their status through careful out-of-sample validations. These examples also clearly show that validation studies benefit from addressing randomized social experiments or large regime shifts (Keane and Wolpin, 2007).[17]

The study of the Bay Area Rapid Transport (BART) is an example of the latter. Before the opening of BART, a regional train service in the San Francisco area, McFadden and associates applied logit models and information about commuters' transport mode choice to predict the use of BART when it became available. After the opening of BART, the commuters were recontacted and actual shares were compared to the predictions. Rather close correspondence between predicted and observed shares was observed.

The PROGRESA experiment refers to a Mexican program to increase schooling levels, by providing substantial payments to parents which are contingent on their children's regular attendance at school. Todd and Wolpin (2006) established a behavioral model of family decisions about fertility and schooling without using post-program data on treated households, when villages were randomly assigned to treatment and control groups. Validations are obtained by comparing predictions about program impacts to those estimated directly from the experiment, and it is concluded that the model produced reasonable forecasts of the effect of the program on school attendance rates of children. As a further validation, treatment effects are derived for the non-treated households too, and compared to outcomes for the treated.

In another validation study, Keane and Wolpin (2007) suggest validating their dynamic programming model of life-cycle decisions of young women by a so-called non-random holdout sample, a sample that differs significantly from the estimation sample along the policy dimension that the model is meant to forecast. Thus, this method differ from a more standard cross-validation, which relies on random holdout samples, as the sample face policy regimes well outside the support of the data. Keane and Wolpin conclude that the model performs well on this and other types of model validations.

Concerning validations of the (static) discrete choice labor supply model, both Blundell (2006) and Brewer et al. (2006) use experimental evidence to qualify their prediction models,[18] whereas

[17] Laboratory experiments can (of course) also be used to validate economic models, see Bajari and Hortacsu (2005).

[18] See also Cai et al. (2008), Hansen and Liu (2011) and Pronzato (2012).

Blundell, Brewer and Francesconi (2008) use a sequence of reforms to discuss the assumption of hours of work flexibility of the labor supply model. Further, Blundell (2006) simulates the effect of the Working Families' Tax Credit (WFTC) reform in the UK, and use a matching differences-in-differences technique (comparison of outcomes for single women eligible to the support with single women not eligible) to obtain results for validation of the model. Then significance tests on the differences between results of the two methods are calculated to qualify that the model predictions do not deviate too much from the experimental evidence. Similar to Blundell (2006), Brewer et al. (2006) denote the difficulties involved when using results of ex-post studies to validate the structural model predictions of effects of the WFTC. Ex-post validations reflect combined impact of the WFTC and contemporaneous changes in the tax and benefit and welfare systems affecting families with children. Correspondingly, one finds substantial variation in the estimates of the WFTC reform in different studies using experimental design, most likely because of differences between studies in the choice of time periods, specifications, etc. Another example of validation of the discrete choice labor supply model is seen in Dagsvik et al. (2014), where the performance of the model is assessed by replications of out-of-sample income distributions.

4. Data sources and introductory data descriptions

Before we in the next section probe deeper into the validation of the structural model, we shall in this section provide some preliminary descriptive evidence, given the tax reform used for identifying the NTR response estimates. We therefore first present the Norwegian tax reform of 2006, and then describe the data sources that we have used in this study, together with some introductory statistics.

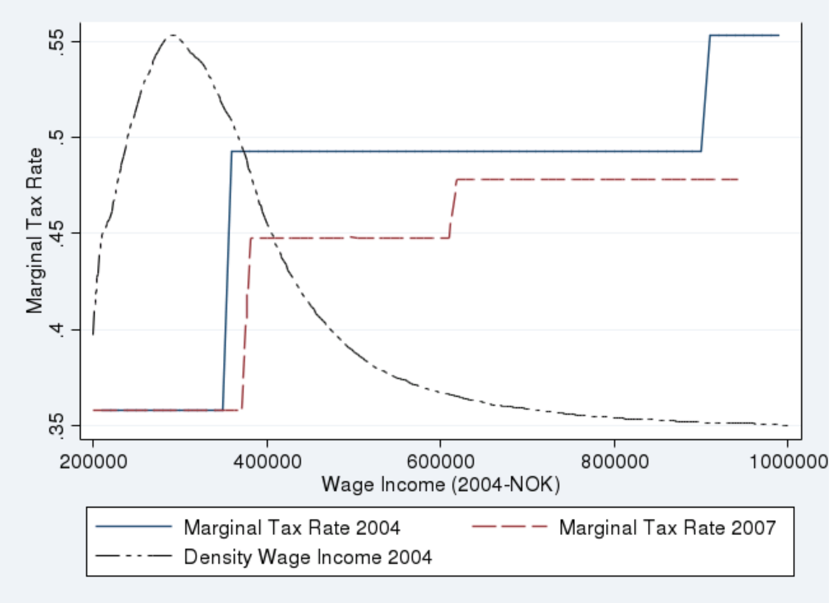
4.1 Reductions in marginal tax rates as a result of the tax reform of 2006

Norway has a "dual income tax" system, enacted through the 1992 tax reform, which consists of a combination of a low proportional tax rate on capital income and progressive tax rates on labor income. The system proliferated in the Nordic countries in the early 1990s. The Norwegian version had a flat 28 percent tax rate levied on corporate income, capital income and labor income coupled with a progressive surtax applicable to labor income. The gap between marginal tax rates on capital income and wage income was problematic, and the schedule was reformed in 2006 in order to narrow the differences, by introducing shareholder income tax, and, most importantly in the present context, by cutting marginal tax rates on labor income.

The tax reform was gradually implemented in 2005 and 2006; in Figure 1 we compare schedules for 2004 (pre-reform) and 2007 (post-reform).[19] The figure shows the principal features of the Norwegian labor income tax system: a two-tier surtax that supplements a basic income tax rate of 28 percent plus a 7.8 percent social security contribution. In 2004, the first tier of the surtax was applied to incomes above NOK354,300 ($52,600/€42,300) at a rate of 13.5 percent, and the second tier of 19.5 percent applied to income in excess of NOK906,900 ($134,600/€108,400). The reform meant that the maximum marginal tax rate fell from 55.3 to 47.8 percent, but became effective at a lower level. In 2007 this threshold was 620,000 ($92,000/€74,000), when recalculated into comparable 2004-values.

It is crucial for identification in the NTR approach that individuals are differently affected by the tax reform. The reform provides a promising schedule for isolating the tax responses, as there is not a monotone relation between initial income level and tax treatment. Moreover, two different tax classes and some regional differences in the tax rates contribute to variations in treatment, independent of initial income levels.

Figure 1. Reductions in marginal tax rates as a result of the tax reform



4.2 Data

We estimate both the discrete choice labor supply model and the equations of the NTR approach by using a large panel data set, Wage Statistics (Statistics Norway, 2006), which is based on administrative registration of employers' reporting of working hours and monthly wages of their

[19] There were some minor adjustments in the schedule from 2006 to 2007 too, which explains why we present the 2007-schedule in Figure 1.

employees. The statistics are collected from a stratified sample of Norwegian firms with at least 3–5 employees (depending on industry). The statistics cover 50–60 percent of the employees in the private sector and 100 percent in public sector. In total, we have information on about 70 percent of Norwegian wage earners. The large number of employees included each year implies that we can utilize the panel dimension of the data source. As we use the same data set for estimation of the structural model and the (panel data) NTR equation, the difference in representativity between private and public sectors is not critical. To estimate the structural model, we use a cross-sectional sample of a pre-reform year; 2004 is chosen here. For the NTR approach we exploit the panel dimension to establish a dataset consisting of overlapping three-year differences (suggested by Feldstein, 1995), over the period 2000–2008.

Information about annual incomes and tax return, family composition, number of children, education, etc. is obtained from the Income Statistics for Persons and Families (Statistics Norway, 2005) and linked to the Wage Statistics, using unique personal identification numbers. Unemployed, self-employed, disabled persons and students are not included in the sample. However, potential wage earners who have chosen not to work are included in the structural model by drawing observations from a sample of non-working individuals, obtained from the Income Statistics for Persons and Families, to match up with the sample of the Wage Statistics. We further limit the sample to persons aged between 25 and 62 years, and we define a person as non-participating if he or she works less than one hour per week.

A main variable of the Wage Statistics is contractual working hours, as reported by the employers. However, in order to approach a measure for actual working hours, we add imputed overtime hours to the contractual working hours. Measures for overtime hours are obtained by dividing monthly overtime payment by individual contractual hourly wage payment, where the latter is calculated by dividing the contractual wage payment by monthly contractual working hours.[20] With respect to income, information on the yearly labor income from the Income Statistics for Persons and Families is used.

Summary statistics of the main variables are presented in Table 1 for two periods, pre-reform (2000–2004) and post-reform (2005–2008), based on approximately 1 million observations each year, over the time period from 2000 to 2008, see also Table B.1 in Appendix B. Table 1 shows that there is little change in average working hours from the first to the second time period.

[20] This approximization may overestimate the overtime hours if overtime payments are higher than compensations for contractual hours.

Table 1. Summary statistics for main variables, cross-sections 2000–2008, individuals aged 25–62. Measures of income and wage rate in Norwegian kroner (NOK)

	2000–2004		2005–2008	
	Mean	Std dev	Mean	Std dev
Contractual working hours	34.0	(7.2)	33.9	(7.4)
Monthly contractual wage income	23,064	(10,012)	27,997	(12,517)
Hourly wage rate (imputed)	159	(54.4)	193	(69.0)
Monthly overtime payment	820	(2,261)	1,040	(2,706)
Total working hours (imputed)	35.2	(8.3)	35.3	(8.5)
Yearly labor income	328,054	(153,708)	406,219	(210,393)

Note: 1 US dollar = 6.74NOK, 1 Euro = 8.37 NOK

4.3 Introductory cross-sectional evidence in an experimental perspective

Can we see any signs of the expected effects of the tax reform in plain data descriptions? Before discussing the results of the NTR panel data approach, we search for signs of the expected responses to the tax reform in data, using the repeated cross-sections of the data material.

As already discussed, the changes of the 2006-reform in marginal tax rates primarily came from changes in the surtax schedule which kicks in approximately at the 66th percentile. We therefore compare average values in two groups of individuals: the 33–66 percentile group and the 66–100 percentile group. Individuals with low incomes, incomes below the 33th percentile, are excluded because they are less suitable for being used in a control group.[21] We have calculated net-of-tax rates, before and after the reform in the two groups, and aligned them with averages for key outcome variables: working hours, earned income and virtual income. In the presentation of results in Table 2, we present normalized values for both earned income and virtual income, letting the sample average for each year be normalized to 1.

Table 2 confirms that individuals at higher income levels experienced a larger increase in the net-of-tax rate than tax-payers at lower income levels. We see some indications of responses in earnings and workings hours when matching up with the differences in changes in the net-of-tax rate, but no clear effects of the tax reform are observed by this simplified approach.[22] There is a small average reduction in working hours (constant average earned income) among individuals in percentile 33–66, indicating a counterfactual (no reform) trend downward in working hours, whereas for individuals in percentile 66-100 (who are believed to be more affected by the lower marginal tax rates) working hours are constant (labor income increase).

[21] The split at the exact 33th percentile is arbitrary and we discuss the robustness of this choice in Appendix B.

[22] More detailed examination of yearly responses indicates that there is a trend towards higher income growth for high income earners, with no visible change exact at the point in time where the reform sets in. Moreover, we do not observe individuals experiencing the largest net-of-tax rates changes by the reform (in the lower part of the surtax range) responding more strongly, compared to the others in surtax position. We therefore need a more sophisticated approach to distinguish between tax responses and trends in the distribution of working hours and earnings, as provided by the NTR literature.

Table 2. Average values for main variables of the NTR approach for different income groups, before and after the 2006 tax reform. Standard deviations in parentheses

	Income percentile 33–66, pre-reform (2000–2004)	Income percentile 33–66, post-reform (2006–2008)	Income percentile 66–100, pre-reform (2000–2004)	Income percentile 66–100, post-reform (2006–2008)
Net-of-tax rate (1-τ)	0.618	0.631	0.510	0.547
	(0.050)	(0.030)	(0.022)	(0.018)
Working hours	36.59	36.29	38.94	38.94
	(5.96)	(6.31)	(6.09)	(6.29)
Normalized earned income	0.98	0.98	1.55	1.61
	(0.08)	(0.08)	(0.57)	(0.70)
Normalized virtual income	0.66	0.69	0.83	0.85
	(0.84)	(0.66)	(1.01)	(0.69)

5. Validations

We now move on to show how simulation results from a discrete choice labor supply can be turned into NTR estimates and compared to the panel data NTR results. First, NTR estimates for working hours and earned income are shown. Then standard elasticity estimates derived from the structural labor supply model are described, before carrying out the actual validation, based on labor supply model simulation results being converted into NTR measures.

5.1 NTR results for working hours and earned income

In the following we closely follow the conventional panel data approach in the NTR literature; see, e.g., Gruber and Saez (2002). We stack observations for each three-year difference (2000–2003, 2001–2004,…, 2005–2008) over the period 2000–2008, and add time invariant explanatory variables as possible explanations for income growth.[23] In Appendix B we show results for alternative time spans.

The estimated equations for working hours and earned income are basically identical. Here, we present the earned income version. The equation for hours of work can simply be obtained by replacing the dependent variable with growth in working hours. When reformulating Equation (2.3) (see Section 2.2) and adding individual control variables, we have that three-year differences in (log) labor income, q_{it}, is explained by a period-specific effect, κ_t, differences in (log) net-of-tax rate, $1-\tau_{it}$, and a set of individual control variables, x_{it},

$$(5.1) \qquad \log\left(\frac{q_{it+3}}{q_{it}}\right) = \kappa_t + \lambda_1 \log\left(\frac{1-\tau_{it+3}}{1-\tau_{it}}\right) + x_{it}\omega + \xi_{it}\,.$$

[23] In order to allow new tax prices to be absorbed by the agents, as already seen, it has become standard to use three-year span in data from pre-reform to post-reform. We will return to a discussion of timing of responses in Section 5.4.

The actual marginal tax rate is not immediately available in the data set, but is constructed by a tax simulation, where incomes are increased by a small amount (five percent). The change in marginal tax rate is clearly endogenous, since the marginal tax rate (as a function of income) is jointly determined with income. In the identification of λ, similar to several other NTR studies, the tax rate change, $\log\left[(1-\tau_{it+3}(q_{it+3}))/(1-\tau_{it}(q_{it}))\right]$, is therefore instrumented by a tax rate change for a "constant" or inflation-adjusted initial income level, $\log\left[(1-\tau_{it+3}((1+b)q_{it}))/(1-\tau_{it}(q_{it}))\right]$, where b corresponds to median income growth from period t to period $t+3$.

The error term in equation (5.1) is correlated with first period working hours, q_{it}, for instance because of mean reversion and drifts in the income distribution (Moffitt and Wilhelm, 2000). Mean reversion stems from individuals with a high number of income in period t, and therefore (mistakenly) placed in the treatment group with large reductions in marginal tax rates, will return to their normal income levels in period $t+3$, and an reduction in income will be recorded. To account for the mean reversion bias, Auten and Carroll (1999) suggest adding $\log q_{it}$ as an additional control variable. As shown in many analyses, Aarbu and Thoresen (2001) included, this control has substantial influence on tax elasticity estimates, and it may shift estimates of the change in the net-of-tax rate from negative to positive. Gruber and Saez (2002) suggest extending the base period income control technique by including a piecewise linear function of $\log q_{it}$.[24] A similar approach is adapted here by using a polynomial in first year's income, but we also show results for the linear mean reversion control (as in Auten and Carroll, 1999).

In the case of working hours, the mean reversion issue might be somewhat less severe, but it is still clearly visible and should not be ignored. We therefore apply the same techniques as for income, but base the control on first year's working hours instead of income.

A main problem of employing rich controls for mean reversion based on first-period information is that identification of the effect of the net-of-tax rate may become blurred, because the mean reversion control and the tax change instrument depend on the same variable; see, for instance, Saez, Slemrod, and Giertz (2012). The problem is alleviated by including periods both with and without tax changes. The identification also benefits from having other sources of variation in the tax rate than income alone: two tax classes (joint and individual taxation) and a separate rate schedule for people in northern Norway are helpful in this respect.

The polynomial function in the log of first period income is not just a control for mean reversion effects; it can also be seen as accounting for changes in the income distribution. For example, a trend

[24] See also Moffitt and Wilhelm (2000) and Kopczuk (2005) on methods to account for mean reversion effects in this type of studies.

towards increasing inequality in income or working hours may result in a spurious correlation between lowered tax rates for high-income individuals and income growth rates.

The Norwegian tax reform of 2006 reduced the tax advantages enjoyed by capital income compared to labor income, and it could therefore result in an income shifting effect where individuals increase their labor earnings at the expense of capital income; see Thoresen and Alstadsæter (2010) for the measurement of income shifting when incentives worked in the opposite direction, in the period prior to the reform.[25] We assume, however, that income shifting is less important in the present context, as responses only for employees are considered.

Income effects are often neglected in the NTR literature, under the assumption that they are close to zero, as found in Gruber and Saez (2002). Moreover, there is no standardized method of constructing controls for the income effect. In our specifications, we have relied on a method proposed by Blomquist and Selin (2010) to approach virtual income (see Section 2). However, there is a collinearity problem, as the two excluded instruments for net-of-tax rate and virtual income are similarly constructed, in particular when categorizing into homogenous groups of individuals. We have therefore decided to omit the representation of income effects in the specifications, under the assumption that we approach the uncompensated effects without them; i.e., they are small, which our preliminary estimations seem to support, see the results reported in Table B.4 in Appendix B.[26]

Individual characteristics are included to control for non-tax-related working hours and income evolution over time or over the lifecycle. x_{it} includes both variables that change over time, and time-invariant variables whose relationship to income may have changed over time. We have had access to a number of socio-demographic characteristics, such as age, years of education, field of education, marital status, number of children, geographical location, and area of origin.

Table 3 shows the results of the 2SLS estimations of Equation (5.1) for both working hours and yearly labor income; Table B.1 in Appendix B provides descriptive statistics for the variables used in the estimation. Results are presented for different specifications, adding in additional control variables sequentially, clearly illustrating the importance of controlling for mean reversion effects. Estimates from the preferred specification, with polynomials of base year working hours/income included, show that the overall uncompensated elasticities for working hours and earnings with respect to the net-of-tax rate are 0.038 and 0.055, respectively. Thus, as the estimated effect for the earned income elasticity exceeds the estimate for working hours, we cannot (at this stage) rule out that the tax-payers have responded to the lower tax rates along other dimensions (as obtaining a higher wage), picked up by the measure for earned income.

[25] See also Gordon and Slemrod (2000) on income shifting.

[26] Thus, we estimate the uncompensated elasticities. Measures of compensated elasticities are rare also in the discrete choice structural labor supply literature; see, however, Dagsvik and Karlström (2005) for a method of obtaining compensated effects.

The estimated (average) net-of-tax elasticities are small compared to most other NTR studies. According to Saez, Slemrod and Giertz (2012), estimates of the elasticity of taxable income from the U.S. (after Feldstein, 1995) range from 0.12 to 0.40. Our estimates, however, measure the responses in working hours and wage earnings only, and will most likely show less responsiveness, compared to estimates for gross income or taxable income (which includes more response dimensions). The estimates are in line with Kleven and Schulz (2014), who report elasticities of approximately 0.05 for wage earners in Denmark.[27]

With respect to working hours, there is an understanding in the literature that the intensive margin responsiveness, which is the main focus here, are modest and sometimes equal to zero (Saez, 2010; Saez, Slemrod and Giertz (2012); Chetty, 2012).[28] It is important to keep in mind that the estimated elasticities reflect average treatment effects of the treated, and will therefore differ dependent on the reform utilized to obtain identification.

Table 3. Estimates of the net-of-tax rate elasticity for working hours and earned income. 2SLS regression results for all wage earners, standard errors in parentheses

	Net-of-tax rate elasticity, working hours	Net-of-tax rate elasticity, earned income
No controls	0.0214***	-0.1878***
	(0.0025)	(0.0028)
Add socioeconomic characteristics	-0.0017	-0.0090***
	(0.0025)	(0.0020)
Add log base year hours/income	0.0481***	0.0221***
	(0.0024)	(0.0020)
Add polynomial of base year hours/income	0.0380***	0.0548***
	(0.0024)	(0.0022)
Number of observations	2,353,603	

Note: Socioeconomic characteristics include gender, wealth, age, age squared, married, number of children under and above the age of 6, newborn, residence in Oslo/ densely populated area, non-western origin, years of education and 9 dummies for field of education. Linear or polynomial control for base year working hours/labor income is included to account for mean reversion. All regressions include year dummies.

Further, we divide the sample into four groups (single females, single males, females in couple, and males in couple), as response estimates for specific groups facilitate closer comparison with the simulation results from the discrete choice model. A third degree polynomial is used as a mean reversion control in the estimations for separate groups.

The results reported in Table 4 suggest that the responses are positive but small for all four groups of wage earners, statistically significant in the range from 0.02 to 0.05. We see the standard pattern of higher elasticities for females (and in particular for females in couples) for the hours of

[27] Singleton (2011) finds earned income responses in the US above this level, in the range from 0.22 to 0.3.

[28] Chetty (2012) explains the small responses in both income and working hours as resulting from optimization errors, an issue we will return to in Section 5.4.

work estimations, whereas the labor income elasticities are more similar in magnitude across the four groups. Whereas the estimates for females do not substantiate that income is more responsive than hours of work, the expected larger earned income elasticity is observed for males, although not strictly significant for single males. This suggests that males might respond along other margins than working hours only, and we can therefore not rule out that there are effects on wages even in a short term perspective.[29]

Table 4. Estimates of the net-of-tax rate elasticity for working hours and earned income. 2SLS regression results for groups of wage earners

	Working hours		Earned income		
	Net-of-tax rate elasticticty	Std error	Net-of-tax rate elasticticty	Std error	Number of observations
Single females	0.0324***	(0.0059)	0.0204***	(0.0051)	353,905
Single males	0.0227***	(0.0055)	0.0392***	(0.0054)	450,519
Females, couple	0.0514***	(0.0046)	0.0312***	(0.0045)	680,881
Males, couple	0.0160***	(0.0037)	0.0525***	(0.0034)	1,162,743

Note: All regressions include control variables for wealth, age, age squared, married, number of children under and above the age of 6, newborn, residence in Oslo/ densely populated area, non-western origin, years of education, 9 dummies for field of education and year dummies. Polynomials of base year working hours or labor income respectively are used as control for mean reversion.

5.2 Standard labor supply model simulations

In the following we shall see how the simulation results of the model can be converted into comparable NTR measures. However, before discussing the results of the validation procedure we present standard wage elasticity estimates, where uncompensated wage elasticities are obtained by increasing the (exogenous) gross hourly wage by one percent and using the model and parameter estimates to simulate the percentage change in predicted hours worked for each individual.

The average elasticity estimates for each population group are shown in Table 5. The wage elasticity is further decomposed into a participation elasticity and an elasticity conditional on participation, measuring the extensive and intensive margin, respectively. As already noted, the results for the intensive margin are more relevant when using the NTR approach for validation.[30] They are small for males, 0.05 and 0.03, and larger for females, 0.17 and 0.24.

[29] Blomquist and Selin (2010) use direct information on wages in a NTR setting and find larger responses, although for a much wider time period (10 years), whereas Blundell, Brewer and Francesconi (2008) find no significant wage effects for single women in response to a sequence of reforms in Britain in the 1990s.

[30] In the conventional NTR framework, described in Section 5.1, the focus is on intensive margin responses. We therefore focus on intensive margin responses in the validation exercise that follows.

Table 5. Wage elasticity estimates derived from simulation of labor supply model, standard errors in parentheses

	Total wage elasticity	Extensive margin wage elasticity	Intensive margin wage elasticity
Single females	0.40 (0.0019)	0.22 (0.0066)	0.17 (0.0055)
Single males	0.29 (0.0089)	0.25 (0.0073)	0.05 (0.0023)
Females in couple	0.46 (0.0182)	0.22 (0.0181)	0.24 (0.0002)
Males in couple	0.06 (0.0264)	0.03 (0.0187)	0.03 (0.0128)

Note: Standard errors obtained by non-parametric bootstrapping, 30 repetitions.

In order to examine to what extent the intensive margin wage elasticity differ over the income distribution, elasticity estimates are derived when the samples have been divided into deciles, based on hourly wage rate rankings; thus, highlighting the nonlinearity characteristic of the model, see Table 6. We see that for the highest deciles, the responses are relatively small, and the response differences between the different groups of tax-payers are smaller, compared to the differences in average measures. Females in the tenth decile are more similar to their high-income male counterparts, rather than to females in other deciles. This implies that the (converted) net-of-tax rate elasticities are not so different across gender (as we soon will show).

Table 6. Intensive margin wage elasticity estimates by wage decile, derived from simulation of labor supply model, standard errors in parentheses

	Single females	Single males	Females in couple	Males in couple
1st decile	0.25 (0.0174)	0.05 (0.0035)	0.24 (0.0004)	0.03 (0.0129)
2nd decile	0.20 (0.0121)	0.06 (0.0030)	0.30 (0.0002)	0.03 (0.0129)
3rd decile	0.19 (0.0084)	0.06 (0.0028)	0.27 (0.0002)	0.03 (0.0133)
4th decile	0.19 (0.0063)	0.05 (0.0025)	0.27 (0.0002)	0.02 (0.0137)
5th decile	0.18 (0.0053)	0.06 (0.0025)	0.26 (0.0002)	0.02 (0.0135)
6th decile	0.19 (0.0044)	0.02 (0.0021)	0.27 (0.0002)	0.02 (0.0133)
7th decile	0.19 (0.0036)	0.02 (0.0014)	0.26 (0.0002)	0.03 (0.0133)
8th decile	0.15 (0.0029)	0.04 (0.0016)	0.23 (0.0001)	0.03 (0.0135)
9th decile	0.12 (0.0024)	0.06 (0.0023)	0.20 (0.0001)	0.04 (0.0137)
10th decile	0.08 (0.0017)	0.08 (0.0030)	0.12 (0.0001)	0.04 (0.0140)

Note: Standard errors obtained by non-parametric bootstrapping, 30 repetitions.

5.3 Converting results from labor supply model simulations into NTR estimates

Next, we show how we can derive estimates of (comparable) net-of-tax rate elasticities from a labor supply model simulation of working hours. As the random utility framework of the discrete choice model implies that a probability distribution for different working time options is generated. In contrast, response estimates found in the NTR literature are derived from marginal optimization, the response estimates (somewhat simplified) reflecting average responses of the “treated”, compared to "the less or not treated".[31] To approach comparable measures, we therefore let the results of labor supply model simulations enter into a regression, similar to that seen in the NTR literature. First, the structural model is used to simulate the pre-reform and post-reform working hours for the four groups of wage earners, see Table 7. Then these results are turned into measures of growth in (simulated) working hours.[32] In the replication of the NTR technology, the variable for the change in the net-of-tax is derived from predicted income levels (hourly wage rate multiplied with predicted hours), and instrumented using similar methods as in the NTR literature: the change in the net-of-tax rate for constant (predicted) pre-reform labor income.

Table 7. Average weekly hours of work, pre- and post-reform, derived from simulation of labor supply model, standard errors in parentheses

	Pre-reform working hours	Post-reform working hours	Difference
Single females	35.20 (0.321)	35.27 (0.322)	0.18 %
Single males	38.95 (0.039)	38.97 (0.040)	0.04 %
Females in couple	32.13 (0.068)	32.25 (0.068)	0.36 %
Males in couple	38.60 (0.013)	38.64 (0.014)	0.11 %

Note: Standard errors obtained by non-parametric bootstrapping, 30 repetitions.

The NTR version of results from the discrete choice labor supply model simulation of effects of the 2006 tax reform are presented in Table 8, see first columns. These NTR estimates are small too, from about 0.02 to about 0.06, with the largest responses seen for single males.

Moreover, in Table 8, the NTR estimates from the labor supply simulations are brought together with the results of the standard NTR evaluation of the reform; the latter have already been presented in Table 4. We see that the panel data NTR measures for working hours are close to the NTR measures obtained from the model simulations. In fact, there is no significant difference between the overall average estimates, see the last row of Table 8. All estimates (for all four groups) are found in the range from 0.02 to 0.06. A difference of 0.04, which is the maximum difference observed for working hours in Table 8 (single males), must be characterized as miniscule, both compared to the variation of elasticity estimates in the literature, see for example the review in Blundell and MaCurdy (1999), and from a policy prediction perspective.[33]

[31] In this perspective the modeling of the NTR literature is therefore more related to the perspective of continuous hours structural labor supply models, such as the so-called Hausman model, see Section 2.2.

[32] Working hours follow from the individual's probability distribution, using a draw from a uniform distribution (the same draw applies for each individual pre- and post-reform). An alternative is to use the expected working hours estimates for each individual pre- and post-reform. This leads to similar results, although the income distribution becomes more compressed by the latter procedure. As in the panel data analysis, the regression is restricted to individuals with predicted pre-reform income in percentile 33 or above.

[33] The revenue effect of erroneously using 0.06 instead of 0.02 can be illustrated by a simple "back-of-the-envelope" calculation for a hypothetical tax change for all wage earners. For a 1 percent change in the net-of-tax rate, when the after-tax additional income growth, due to 0.06 instead of 0.02, is multiplied by the number of people in the group (2,739,000), the total effect will not exceed 100 million NOK (compared to a total revenue from the income tax for persons of around 280 billion NOK in 2014).

Thus, the model performs well according to this validation. Of course, this does not mean that the simulation model is approved; it only implies, according to our judgment, that the model has not been rejected by the present test.

The largest deviations between the response estimates of the labor supply model and the traditional NTR estimates on working hours are observed for single males and females in couples, suggesting that the labor supply model overstates the responses for single males and underrates the responses for married and cohabiting females.

Further, the earned income responses are on average somewhat larger, which suggest that it might be important from a forecasting perspective to bring in effects from changes in wages in the simulation model. As already noted, combining the simulation model with other modeling tools, as a general equilibrium model, is one option. To account for a (positive) tie between hours and wages directly in the simulations appears to be a complex empirical challenge.[34]

Table 8. Comparison of net-of-tax rate elasticity estimates obtained from labor supply model simulations and the NTR approach for working hours and earned income. Standard errors in parentheses

	Discrete choice labor supply model simulations, working hours	Panel data information	
		Working hours	Earned income
Single females	0.018 (0.0005)	0.032 (0.0037)	0.020 (0.0051)
Single males	0.062 (0.0027)	0.023 (0.0055)	0.039 (0.0054)
Females in couple	0.026 (0.0001)	0.051 (0.0046)	0.031 (0.0045)
Males in couple	0.015 (0.0005)	0.016 (0.0059)	0.053 (0.0034)
Weighted average	0.026 (0.0012)	0.028 (0.0053)	0.041 (0.0043)

Note: The weighted averages are calculated by accounting for the number of observations in each group. Standard errors are obtained by using the so-called delta method.

5.4 Time dependency and response frictions

The preceding discussion more generally points to interpretational challenges in the present validation, i.e., which effects are in reality picked up by the NTR estimates? For example, one may question the time span used in the NTR approach. Both approaches to obtain tax responses can be criticized for not accounting for key elements of the optimization process, as adjustments costs and inattention (Chetty et al., 2011; Chetty, 2012). Under such characteristics of the optimization, one may question how long it takes for the tax-payer to be established in a new optimum. Jia and Vattø (2014) find that the labor supply model responses are considerably more sluggish when allowing for state dependence and adjustment costs. For females in couples, only about one third of the full effect is reached in the first

[34] Aaronson and French (2009) provide a theoretical argument for why it is plausible to expect a positive relationship between offered wage rates an hours of work (after controlling for selection effects in the measurement of wages).

year of a policy change, and close to the full effect is reached after about 7 years. Given this, the standard three year time span of the NTR literature may be too short in the current context.[35] The results presented in Table B.3 of Appendix B indicate some time span dependency of the NTR estimates. For example, we see larger responses for working hours when extending the measurement time period from three to four years.

Moreover, given that there are adjustments costs, reforms must be large enough to overcome the frictions for effects to materialize in data (Chetty, 2012). Figure 1 shows that the 2006-reform does not involve very large changes in incentives.[36] The variation in the net-of-tax rate changes over the income range in the treatment group means that we may observe more responsiveness in the group experiencing the largest change (15.5 percent). However, we find no signs of such effects. In extension of this, we have also calculated bounds for the true structural parameter, along the lines of Chetty (2012). Given our relatively small changes in tax rates, which will result in rather wide bounds according to the procedure of Chetty, we cannot rule out that our NTR estimates have been substantially attenuated by optimization frictions, i.e., the reform is not large enough to obtain accurate estimates.

6. Conclusion

The discrete choice labor supply model is a tool that is frequently used to analyze a wide range of hypothetical tax and benefit reforms. Given its key role in the decision-making process, it is important to validate its capacity to provide reasonable descriptions of the effects of prospective policies. There has recently been growing interest in validating discrete choice structural models using natural experiments. However, we have yet to see any detailed discussion of how the standard structural labor supply model can be validated by using methods from the NTR literature.

A validation that is simply based on comparisons of average wage elasticities from the labor supply model with average net-of-tax rates from the NTR approach is misleading. The reason is that NTR estimates are derived from specific tax reforms, and therefore measure the average effects for the treated individuals. The nonlinearity of the discrete choice labor supply model, on the other hand, implies different responses along the income distribution.

In this study, we have shown how a standard discrete choice labor supply model, similar to one made available to Norwegian decision-makers through the model system LOTTE, is validated by NTR estimates of working hours and earnings. The estimated structural model is used to simulate the labor supply effects of the Norwegian tax reform of 2006. Working hours are simulated pre- and post-

[35] The structure of overlapping panels also contribute to underestimate the long run effect if behavioral responses take more than one year, see Bækgaard (2014).

[36] At least not compared to the changes reported in Table 1 in Chetty (2012). See also Bastani and Selin (2014) for analysis of a large Swedish reform.

reform under an exogenous wage assumption, and the regression framework of the NTR literature is used to obtain net-of-tax rate elasticities. These estimates have then been compared with NTR estimates obtained in the conventional manner.

Our main finding is that simulations from the structural labor supply model yield net-of-tax elasticity estimates that are close to the elasticities estimated on basis of the panel data. Thus, we find it reassuring that the predictions of the labor supply model are not far from the results of the alternative framework. Both approaches point to very modest effect of the reform.

As we validate the labor supply model with respect to panel data information both on working hours and earned income, we have also discussed the implication of ignoring responses in wages when describing policy effects. Even though we do not see a clear picture, on average the earned income responses appear to be somewhat larger than the effects on working hours alone, which may have resulted from responses in wages too. Further developments of the discrete choice model to fit with key response margins are certainly appreciated, though empirically challenging.

Finally, our results give support to using the NTR approach for earned income to validate the structural model, when information on working hours is absent or insufficient.

References

Aaberge, R., U. Colombino, and S. Strøm (1999). Labour Supply in Italy: An Empirical Analysis of Joint Household Decisions, with Taxes and Quantity Constraints, *Journal of Applied Econometrics*, 14, 403–422.

Aaberge, R., J.K. Dagsvik, and S. Strøm (1995). Labor Supply Responses and Welfare Effects of Tax Reforms. *Scandinavian Journal of Economics*, 97, 635–659.

Aarbu, K.O. and T.O. Thoresen (2001). Income Responses to Tax Changes – Evidence from the Norwegian Tax Reform. *National Tax Journal*, 54, 319–35.

Aaronson, D. and E. French (2009). The Effect of Progressive Taxation on Labor Supply with Hours and Wages are Jointly Determined. *Journal of Human Resources*, 44, 386–408.

Aasness, J., J. K. Dagsvik, and T. O. Thoresen (2007). "The Norwegian Tax-benefit Model System LOTTE". In A. Gupta and A. Harding (eds.): *Modelling Our Future: Population Ageing, Health and Aged Care, International Symposia in Economic Theory and Econometrics,* Amsterdam: Elsevier Science, North-Holland, 513–518.

Angrist, J.D., and J.-S. Pischke (2009). *Mostly Harmless Econometrics: An Empiricists Companion*. Princeton: Princeton University Press.

Angrist, J.D. and J.-S. Pischke (2010). The Credibility Revolution in Empirical Economics: How Better Research Design is Taking the Con out of Econometrics. *Journal of Economic Perspectives*, 24, 3–30.

Auten, G. and R. Carroll (1999). The Effect of Income Taxes on Household Income. *Review of Economics and Statistics*, 81, 681–693.

Bajari, P. and A. Hortacsu (2005). Are Structural Estimates from Auction Models Reasonable? Evidence from Experimental Data. *Journal of Political Economy*, 113, 703–41.

Bastani, S. and H. Selin (2014). Bunching and Non-Bunching at Kink Points of the Swedish Tax Schedule. *Journal of Public Economics*, 109, 36–49.

Bingley, P. and I. Walker (1997). The Labour Supply, Unemployment and Participation of Lone Mothers in In-work Transfer Programmes. *Economic Journal*, 107, 1375–1390.

Blomquist, S. and H. Selin (2010). Hourly Wage Rate and Taxable Labor Income Responsiveness to Changes in Marginal Tax Rates. *Journal of Public Economics*, 94, 878–889.

Bloemen, H.G. and A. Kapteyn (2008). The Estimation of Utility Consistent Labor Supply Models by Means of Simulated Scores. *Journal of Applied Econometrics*, 23, 395–422.

Blundell, R. (2006). Earned Income Tax Credit Policies: Impact and Optimality. The Adam Smith Lecture 2005. *Labour Economics*, 13, 423–443.

Blundell, R., M. Brewer, and M. Francesconi (2008). Job Changes, Hours Changes and the Path of Labour Supply Adjustment. *Journal of Labor Economics*, 26, 421–445.

Blundell, R., A. Duncan, J. McCrae, and C. Meghir (2000). The Labour Market Impact of the Working Families' Tax Credit. *Fiscal Studies*, 21, 75–104.

Blundell, R. and T. MaCurdy (1999). "Labor Supply: A Review of Alternative Approaches". In O.C. Ashenfelter and D. Card (eds.): *Handbook of Labor Economics*, Vol. 3A, Amsterdam: North-Holland, 1559–1695.

Blundell, R. and A. Shephard (2012). Employment, Hours of Work and the Optimal Taxation of Low-Income Families. *Review of Economic Studies*, 79, 481–510.

Brewer, M., A. Duncan, A. Shephard, and M.-J. Suárez (2006). Did Working Families' Tax Credit Work? The Impact of In-Work Support on Labour Supply in Great Britain, *Labour Economics*, 13, 699–720.

Burtless, G. and J. Hausman (1978). The Effect of Taxes on Labor Supply. *Journal of Political Economy*, 86, 1103–1130.

Bækgaard (2014). The Difference-in-Difference Approach with Overlapping Differences – Experimental Verification of Estimation Bias. DREAM working paper 2014:3.

Cai, L., G. Kalb, Y.-P. Tseng, and H. Vu (2008). The Effect of Financial Incentives on Labour Supply: Evidence for Lone Parents from Microsimulation and Quasi-Experimental Evaluation. *Fiscal Studies*, 29, 285–325.

Chetty, R. (2009). Sufficient Statistics for Welfare Analysis: A Bridge Between Structural and Reduced-Form Methods. *Annual Review of Economics*, 1, 451–87.

Chetty, R. (2012). Bounds on Elasticities with Optimization Frictions: A Synthesis of Micro and Macro Evidence on Labor Supply. *Econometrica*, 80, 969–1018.

Chetty, R., J.N. Friedman, T. Olsen, and L. Pistaferri (2011). Adjustment Costs, Firm Responses, and Micro vs. Macro Labor Supply Elasticities: Evidence from Danish Tax Records. *Quarterly Journal of Economics*, 126, 749–804.

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

Creedy, J., G. Kalb, and R. Scutella (2006). Income Distribution in Discrete Hours Behavioural Microsimulation Models: An Illustration. *Journal of Economic Inequality*, 4, 57–76.

Creedy, J. and A. Duncan (2005). Aggrgating Labour Supply and Feedback Effects in Microsimulation. *Australian Journal of Labour Economics*, 8, 277–290.

Dagsvik, J.K. and Z. Jia (2014). Labor Supply as a Choice among Latent Jobs: Unobserved Heterogeneity and Identification, forthcoming *Journal of Applied Econometrics*.

Dagsvik, J.K., Z. Jia, T. Kornstad and T.O. Thoresen (2014). Theoretical and Practical Arguments for Modeling Labor Supply as a Choice among Latent Jobs. *Journal of Economic Surveys*, 28, 134–151.

Dagsvik, J.K. and A. Karlström (2005). Compensating variation and Hicksian choice probabilities in random utility models that are nonlinear in income. *Review of Economic Studies*, 72, 57–76.

Dagsvik, J.K. and S. Strøm (2006). Sectoral Labor Supply, Choice Restrictions and Functional Form. *Journal of Applied Econometrics*, 21, 803–826.

Deaton, A. (2010). Instruments, Randomization, and Learning about Development. *Journal of Economic Literature*, 48, 424–455.

Duncan, A. and C. Giles (1996). Labour Supply Incentives and Recent Family Credit Reforms. *The Economic Journal*, 106, 142–155.

Duncan, A. and M. Weeks (2000). "Transitions Estimators in Discrete Labour Supply Models. In L. Mitton, H. Sutherland and M. Weeks (eds.): *Microsimulation Modelling for Policy Analysis: Challenges and Innovations*, Cambridge: Cambridge University Press, 292–306.

Eissa, N. and H. Hoynes (2006). "Behavioral Responses to Taxes: Lessons from the EITC and Labor Supply." In J.M. Poterba (ed.), *Tax Policy and the Economy*, Volume 20, Cambridge, MA: MIT Press, 74–110.

Eissa, N. and J.B. Liebman (1996). Labor Supply Response to the Earned Income Tax Credit. *Quarterly Journal of Economics*, 111, 605–637.

Feldstein, M. (1995). The Effect of Marginal Tax Rates on Taxable Income: A Panel Study of the 1986 Tax Reform Act. *Journal of Political Economy*, 103, 551–572.

Feldstein, M. (1999). Tax Avoidance and the Deadweight Loss of the Income Tax. *Review of Economics and Statistics*, 81, 674–680.

Gelber, A. (2014). Taxation and the Earnings of Husbands and Wives: Evidence from Sweden, forthcoming in *Review of Economics and Statistics*.

Goolsbee, A. (1999). Evidence on the High-Income Laffer Curve from Six Decades of Tax Reform**.** *Brookings Papers on Economic Activity*, 1999, 1–64.

Gordon, R.H. and J.B. Slemrod (2000). Are “Real” Responses to Taxes Simply Income Shifting Between Corporate and Personal Tax Bases? In J. Slemrod (ed.): *Does Atlas Shrug? The Economic Consequences of Taxing the Rich*, New York: Russell Sage Foundation, 240–280.

Gruber, J. and E. Saez (2002). The Elasticity of Taxable Income: Evidence and Implications. *Journal of Public Economics*, 84, 1–32.

Haan, P. (2006). Much Ado About Nothing: Conditional Logit vs Random Coefficient Models for Estimating Labour Supply Estimates. *Applied Economics Letters*, 13, 251–256.

Haan, P. and V. Steiner (2005). Distributional Effects of the German Tax Reform 2000 – a Behavioral Microsimulation Analysis. *Journal of Applied Social Science Studies*, 125, 39–49.

Hansen, J. and X. Liu (2011). Estimating Labor Supply Responses and Welfare Participation: Using a Natural Experiment to Validate a Structural Labor Supply Model, IZA Discussion Paper No. 5718, Bonn, Germany.

Hausman, J.A. (1985). The Econometrics of Nonlinear Budget Sets. *Econometrica*, 53, 1255–1282.

Heckman, J.J. (1979). Sample Selection Bias as a Specification Error. *Econometrica*, 47, 153–161.

Heckman, J.J. (2010). Building Bridges Between Structural and Program Evaluation Approaches to Evaluating Policy. *Journal of Economic Literature*, 48, 356–398.

Heckman, J.J. and S. Urzua (2010). Comparing IV with Structural Models: What Simple IV Can and Cannot Identify. *Journal of Econometrics*, 156, 27–37.

Heckman, J.J. and E. Vytlacil (2005). Structural Equations, Treatment Effects, and Econometric Policy Evaluation. *Econometrica*, 73, 669–738.

Holmlund, B. and M. Söderström (2011). Estimating Dynamic Income Responses to Tax Reform. *The B.E. Journal of Economic Analysis & Policy*, 11, Iss. 1 (Contributions), Article 71.

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

Jia, Z. and T. Vattø (2014). Tax Response Inertia in Labor Supply: Effects of State Dependence in Preferences and Opportunities, manuscript, Statistics Norway.

Keane, M.P. (2010a). Structural vs Atheoretic Approaches to Econometrics. *Journal of Econometrics*, 156, 3–20.

Keane, M.P. (2010b). A Structural Perspective on the Experimentalist School. *Journal of Economic Perspectives*, 24, 47–58.

Keane, M.P. (2011). Labor Supply and Taxes: A Survey. *Journal of Economic Literature*, 49, 961–1075.

Keane M. and K. Wolpin (2007). Exploring the Usefulness of a Non-Random Holdout Sample for Model Validation: Welfare Effects on Female Behavior. *International Economic Review*, 48, 1351–78.

Kleven, H.J. and E.A. Schulz (2014). Estimating Taxable Income Responses using Danish Tax Reforms. *American Economic Journal: Economic Policy*, forthcoming.

Kniesner, T.J. and J.P. Ziliak (2008). “Evidence of Tax-Induced Individual Behavioral Responses”. In J.W. Diamond and G.R. Zodrow (eds.): *Fundamental Tax Reform: Issue, Choices, and Implication*, Cambridge (MA): MIT Press, 375–411.

Kopczuk, W. (2005). Tax Bases, Tax Rates, and the Elasticity of Reported Income. *Journal of Public Economics*, 89, 2093–2119.

Kornstad, T. and T.O. Thoresen (2006). Effects of Family Policy Reforms in Norway: Results from a Joint Labour Supply and Childcare Choice Microsimulation Analysis. *Fiscal Studies*, 27, 339–371.

Labeaga J.M., X. Oliver, and A. Spadaro (2008). Discrete Choice Models of Labour Supply, Behavioural Microsimulation and the Spanish Tax Reforms. *Journal of Economic Inequality*, 6, 247–273.

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

Lindsey, L. (1987). Estimating the Behavioral Responses of Taxpayers to Changes in Tax Rates: 1982-1984. *Journal of Public Economics*, 33, 173–206.

McFadden, D. (1977). Demand Model Estimation and Validation (with A.P. Talvitie and Associates) Urban Travel Demand Forecasting Project, Final Report, Volume V, Institute of Transportation Studies, University of California, Berkeley, June 1977.

McFadden, D. (1984). “Econometric analysis of qualitative response models”. In Z. Griliches and M.D. Intriligator (eds.): *Handbook of Econometrics*, Vol. 2, Amsterdam: North-Holland, 1385–1457.

McFadden, D. and K. Train (2000). Mixed MNL Models for Discrete Response. *Journal of Applied Econometrics*, 15, 447–470.

Meghir, C. and D. Phillips (2010). “Labour Supply and Taxes”. In J. Mirrlees, S. Adam, T. Besley, R. Blundell, S. Bond, R. Chote, M. Gammie, P. Johnson, G. Myles, and J. Poterba (eds.): *Dimensions of Tax Design: The Mirrlees Review*, Oxford: Oxford University Press.

Moffitt, R. (1984). The Estimation of a Joint Wage Hours Labor Supply Model. *Journal of Labor Economics*, 2, 550–566.

Moffitt, R. and M.O. Wilhelm (2000). “Taxation and the Labor Supply Decisions of the Affluent”. In J. Slemrod (ed.): *Does Atlas Shrug? The Economic Consequences of Taxing the Rich*, New York: Russell Sage Foundation, 193–234.

Peichl, A. and S. Siegloch (2012). Accounting for Labor Demand Effects in Structural Labor Supply Models. *Labour Economics*, 19, 129–138.

Pronzato, C. D. (2012). Comparing Quasi-Experimental Designs and Structural Models for Policy Evaluation: The Case of a Reform of Lone Parental Welfare, IZA Discussion paper series No. 6803, Bonn, Germany.

Saez, E. (2010). Do Taxpayers Bunch at Kink Points? *American Economic Journal: Economic Policy*, 2, 180–212.

Saez, E., J.B. Slemrod, and S.H. Giertz (2012). The Elasticity of Taxable Income with Respect to Marginal Tax Rates: a Critical Review. *Journal of Economic Literature*, 50, 3–50.

Singleton, P. (2011). The Effect of Taxes on Taxable Earnings. Evidence from the 2001 and Related US Federal Tax Acts. *National Tax Journal*, 64, 323–352.

Statistics Norway (2003). Labour force survey 2001, Official Statistics of Norway (NOS D 748)

Statistics Norway (2005). Income Statistics for Persons and Families 2002-2003, Official Statistics of Norway (NOS D 338).

Statistics Norway (2006). Wage statistics 2005. Official Statistics of Norway (NOS D 362).

Thoresen, T.O. and A. Alstadsæter (2010). Shifts in Organizational Form under a Dual Income Tax System. *FinanzArchiv/Public Finance Analysis*, 66, 384–418.

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

Train, K. (2009): *Discrete Choice Methods with Simulation*. Second edition. New York: Cambridge University Press.

UN (2007). Register-Based Statistics in the Nordic Countries. Review of Best Practices with Focus on Population and Social Statistics. United Nations Publications, New York.

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

Van Soest, A., M. Das, and X. Gong (2002). A Structural Labour Supply Model with Flexible Preferences. *Journal of Econometrics*, 107, 345–374.

Vattø, T.E. (2014). The Dynamics of Earnings Responses to Tax Changes, manuscript, Statistics Norway.

Appendix A. Estimation of the discrete choice model

As noted in Section 2.1 the discrete model validated in the present study is a generalized version of the model summarized in section 2, which we have denoted the "job choice model". The model departs in an essential way from previous approaches in its focus on a more comprehensive description of the choice environment in which job choice is the fundamental decision variable. In practice, however, it provides a rationalization for the state-specific dummy variables which are added to the deterministic part of the utility function in the approach of van Soest (1995).

A job is characterized with fixed (job-specific) working hours, wages and other nonpecuniary attributes. We shall assume that the hours of work take only a finite number of values, represented by the set *D*. Further, let $B(h)$ denote the agent's set of available jobs with hours of work *h*. Let $m(h)$ be the number of jobs in $B(h)$, a measure of opportunity (unobserved to the researcher). There is only one nonmarket alternative, so that $m(0) = 1$. When inserting the opportunity measure into the expressions for probabilities, we obtain

$$\text{(A.1)}\qquad P(h)=\frac{\exp v(f(hw,I),h)m(h)}{\exp v(f(hw,I),0)+\sum_{h\in D}\exp v(f(hw,I),h)m(h)}.$$

Equation A.1 yields choice probabilities that are analogous to multinomial logit ones with representative utility terms weighted by the frequencies of available jobs, $\{m(h)\}$. Unfortunately, $\{m(h)\}$ is not directly observable, but under specific assumptions, one can identify $m(h)$ and $\exp v(\cdot)$ and estimate their parameters, see Dagsvik and Jia (2014) and Dagsvik et al. (2014) for further details. From a simplified perspective, we can see this version of the discrete choice model as a standard specification in which we allow for alternative specific constant terms for non-participation and full time work (35–40 hours per week).

The model is estimated for single females, singles males and for females and males in couples. Note that for persons in couples we estimate individual models, when the income of the spouse enters into the budget restriction as non-labor income. The discretization is obtained dividing into 5 categories based on weekly hours of work: For females $h\in\langle 0-1,1-20,20-35,35-40,40+\rangle$ and for males $h\in\langle 0-1,1-35,35-40,40-55,55+\rangle$. Weekly working hours are measured from the sum of contractual working hours and imputed overtime hours from the Wage Statistics, presented in Section 4.2. About 25 percent of the wage earners report positive overtime payment, for the others total working hours equals contractual working hours. We assume that the constructed measure of total working hours per week is a good proxy for a “normal” working week during the year. An alternative is to use contractual hours of work only, but we then loose some of the variation and responses in working hours. As we focus on tax changes at high income levels in the present study, it is important to allow for responses through increased overtime work.

In order to estimate the multinomial logit model,[37] it is necessary to simulate the counterfactual disposable income levels for each discrete alternative, for each individual. We compute the hourly wage as monthly contractual wage income divided by contractual working hours for the same month, as reported in the Wage Statistics. Individuals with improbably low or high computed hourly wage rates (under NOK60 ($8.9/€7.2) or above NOK3,500 ($519/€418) in 2004) were excluded. The log of computed wage rates is then regressed on individual characteristics in a Heckman selection regression (Heckman, 1979), which takes into account that individual unobservable effects influencing the wage and participation in the labor market might be correlated. We find evidence for this by positive and significant Mills lambda parameter for all groups, except single females. For all

[37] This type of multinomial logit model with alternative-varying regressors is also called a conditional logit model; see McFadden (1984).

individuals, across all choices, we used the predicted individual wage rate, accounting for a random effect by adding an error term, based on draws (30 draws per individual) from a normal distribution.

The actual and counterfactual consumption levels are simulated by multiplying the wage rate by the median working hours point of the discrete intervals, $C = f(hw, I)$, where a tax simulation program is used to simulate taxes and disposable income for each individual's hypothetical working hours choice. For couples, income of the spouse is assumed to be exogenously given and included in non-labor income.

Tables A.1 and A.2 report the results of the wage regressions, whereas tables A.3 and A.4 show the results of the estimation of the labor supply model. For all four groups we observe positive marginal utility of both consumption and leisure (α_0 and $\beta_0 + \gamma X$ are positive), and α_1 and β_1 are less than 1, which implies that the likelihood functions are strictly concave.

In order to evaluate the estimation results, Figure A.1 shows diagrams of the actual frequencies of working hours and the corresponding probability distribution based on model simulations, for single females, single males, and females and males in couple. The simulated probabilities are derived by calculating the average probability for each choice of hours, based on the individual probabilities. We see that there is close correspondence between observed and predicted choices.

Table A.1. Estimation results of wage regressions for single females and single males: log of hourly wage as the dependent variable

	Single females		Single males	
	Coefficient	Std error	Coefficient	Std error
Experience	0.0164***	(0.0002)	0.0214***	(0.0004)
Experience squared	-0.0003***	(0.0000)	-0.0003***	(0.0000)
Low education	-0.0987***	(0.0021)	-0.1521***	(0.0045)
High education	0.2520***	(0.0012)	0.3007***	(0.0026)
Residence in densely populated area	0.0672***	(0.0009)	0.0585***	(0.0020)
Non-western origin	-0.0988**	(0.0034)	-0.2657**	(0.0073)
Field of education				
General	-0.0270***	(0.0037)	0.1540***	(0.0086)
Human, art	-0.1073***	(0.0037)	-0.0493***	(0.0090)
Education	-0.1062***	(0.0040)	-0.0087	(0.0103)
Social sc., law	-0.0280***	(0.0043)	0.0681***	(0.0101)
Business, administration	-0.0233***	(0.0037)	0.1383***	(0.0088)
Natural sc., technology	-0.0132***	(0.0039)	0.1366***	(0.0088)
Health	-0.1200***	(0.0038)	0.0301***	(0.0100)
Primary industries	-0.0707***	(0.0061)	0.0359***	(0.0105)
Service	-0.0733***	(0.0045)	0.0914***	(0.0093)
Constant	4.7759***	(0.0049)	4.6084***	(0.0107)
Selection (Participation=1)				
Experience	0.0649***	(0.0020)	0.0093***	(0.0021)
Experience squared	-0.0013***	(0.0000)	-0.0003***	(0.0000)
Low education	-0.3758***	(0.0184)	-0.2693***	(0.0210)
High education	0.3785***	(0.0156)	0.2955***	(0.0158)
Residence in densely populated area	0.0296**	(0.0108)	-0.0840***	(0.0187)
Non-western origin	-1.0045***	(0.0176)	-0.8196***	(0.0110)
Field of education				
General	0.8136***	(0.0220)	0.9336***	(0.0231)
Human, art	0.5469***	(0.0259)	0.6420***	(0.0314)
Education	1.2989***	(0.0343)	1.4583***	(0.0501)
Social, law	0.9963***	(0.0406)	1.0382***	(0.0412)
Business, administration	0.9621***	(0.0224)	1.0231***	(0.0246)
Nature, technology	0.8982***	(0.0257)	1.1249***	(0.0203)
Health	1.2467***	(0.0227)	1.3472***	(0.0389)
Primary industries	0.7325***	(0.0548)	0.8841***	(0.0405)
Service	0.9259***	(0.0365)	1.1395***	(0.0292)
Excluded variables				
Children under age 3	-0.4497***	(0.0244)	-0.4106***	(0.0923)
Children under age 6	-0.3411***	(0.0166)	0.1855**	(0.0569)
Wealth	0.0199***	(0.0037)	0.1120***	(0.0039)
Nonlabor income	-0.0207***	(0.0016)	-0.0625***	(0.0015)
Constant	0.2297***	(0.0282)	0.8837***	(0.0289)
Mills lambda	-0.0187**	(0.0064)	0.3678***	(0.0172)
Number of observations	187,829		168,793	

Note: *significant at 0.10 level, ** significant at 0.05 level, ***significant at 0.01 level

Table A.2. Estimation results of wage regressions for males and females in couple: log of hourly wage as the dependent variable

	Females in couple		Males in couple	
	Coefficient	Std error	Coefficient	Std error
Experience	0.0165***	(0.0002)	0.0253***	(0.0002)
Experience squared	-0.0003***	(0.0000)	-0.0004***	(0.0000)
Low education	-0.1107***	(0.0015)	-0.1800***	(0.0024)
High education	0.2873***	(0.0009)	0.3439***	(0.0013)
Residence in densely populated area	0.0732***	(0.0007)	0.0977***	(0.0011)
Non-western origin	-0.1574**	(0.0023)	-0.2656***	(0.0035)
Field of education				
General	0.0662***	(0.0032)	0.1353***	(0.0057)
Human, art	-0.0172***	(0.0032)	-0.1158***	(0.0060)
Education	-0.0103**	(0.0033)	-0.1649***	(0.0061)
Social sc., law	0.0824***	(0.0037)	0.0102	(0.0061)
Business, administration	0.0723***	(0.0032)	0.1019***	(0.0057)
Natural sc., technology	0.0945***	(0.0033)	0.0488***	(0.0057)
Health	-0.0204***	(0.0033)	-0.0858***	(0.0062)
Primary industries	0.0245***	(0.0050)	-0.0456***	(0.0062)
Service	0.0298***	(0.0038)	0.0048	(0.0059)
Constant	4.6585***	(0.0043)	4.7084***	(0.0065)
Selection (Participation=1)				
Experience	0.0138***	(0.0019)	-0.0202***	(0.0034)
Experience squared	-0.0006***	(0.0000)	0.0002**	(0.0001)
Low education	-0.3263***	(0.0125)	-0.0709**	(0.0248)
High education	0.5611***	(0.0117)	0.2303***	(0.0183)
Residence in densely populated area	-0.0622***	(0.0081)	-0.0535***	(0.0140)
Non-western origin	-0.9761***	(0.0125)	-0.7241***	(0.0188)
Field of education				
General	0.9610***	(0.0162)	0.9477***	(0.0257)
Human, art	0.6226***	(0.0200)	0.7551***	(0.0386)
Education	1.1249***	(0.0228)	1.2556***	(0.0490)
Social, law	0.9849***	(0.0323)	0.8755***	(0.0466)
Business, administration	1.0624***	(0.0165)	0.9709***	(0.0275)
Nature, technology	1.0394***	(0.0196)	1.1481***	(0.0227)
Health	1.2449***	(0.0165)	1.3206***	(0.0519)
Primary industries	0.6506***	(0.0414)	0.7054***	(0.0407)
Service	0.9818***	(0.0274)	1.1346***	(0.0340)
Excluded variables				
Children under age 3	-0.0265*	(0.0110)	0.0044	(0.0197)
Children under age 6	-0.3765***	(0.0082)	-0.0876***	(0.0143)
Wealth	0.0100***	(0.0025)	0.0030	(0.0040)
Partners total income	-0.1899***	(0.0056)	0.2252***	(0.0068)
Constant	1.4467***	(0.0328)	1.0084***	(0.0501)
Mills lambda	0.1303***	(0.0049)	0.2070***	(0.0167)
Number of observations	358,776		307,292	

Note: *significant at 0.10 level, ** significant at 0.05 level, ***significant at 0.01 level

Table A.3. Estimation results for the discrete choice labor supply model. Single females and single males

		Single females		Single males	
		Coefficient	Std error	Coefficient	Std error
Consumption					
Constant (Scale 10^{-4})	α_0	0.4225***	(0.0070)	0.7292***	(0.0076)
Exponent	α_1	0.9244***	(0.0059)	0.6360***	(0.0035)
Leisure					
Age (Scale 1/10)	γ_1	-0.9341***	(0.0726)	-0.1781**	(0.0623)
Age Squared (Scale 1/100)	γ_2	0.1428***	(0.0087)	0.0390***	(0.0074)
# Children under 6 years	γ_3	-0.7002***	(0.0360)	-0.8852***	(0.0442)
# Children above 6 years	γ_4	-0.4812***	(0.0200)	-0.7012***	(0.0242)
Constant (Scale 1/80)	β_0	5.5632***	(0.2096)	3.1620***	(0.1452)
Exponent	β_1	-1.0931***	(0.0425)	-0.7665***	(0.0232)
Alternative specific constants					
Non-participation	f_1	-0.9429***	(0.0267)	1.2026***	(0.0257)
Full-time	f_4/f_3	1.3328***	(0.0083)	0.9853***	(0.0065)
Number of observations		187,165		168,340	

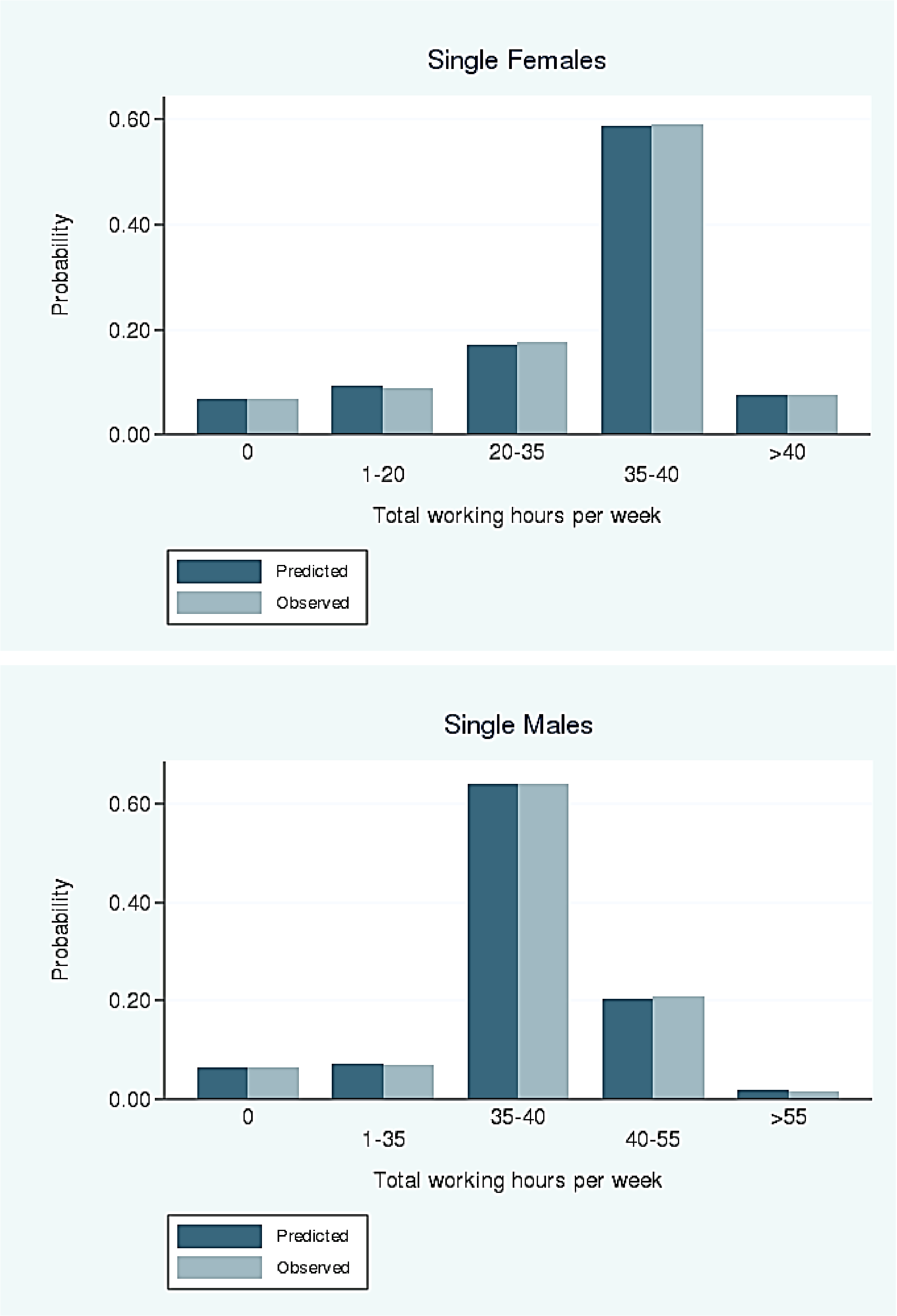
Note: *significant at 0.10 level, ** significant at 0.05 level, ***significant at 0.01 level

Table A.4. Estimation results for the discrete choice labor supply model. Females and males in couple

		Females in couple		Males in couple	
		Coefficient	Std error	Coefficient	Std error
Consumption					
Constant (Scale 10^{-4})	α_0	0.6580***	(0.0071)	0.6728***	(0.0144)
Exponent	α_1	0.9248***	(0.0027)	0.5973***	(0.0073)
Leisure					
Age	γ_1	-0.7688***	(0.0286)	-0.0237***	(0.0058)
Age squared	γ_2	0.1061***	(0.0034)	0.0040***	(0.0008)
# Children under 6 years	γ_3	0.1274***	(0.0045)	0.0007	(0.0006)
# Children above 6 years	γ_4	0.0271***	(0.0030)	-0.0033***	(0.0007)
Constant (Scale 1/80)	β_0	2.8846***	(0.0786)	0.1103***	(0.0197)
Exponent	β_1	-2.9177***	(0.0294)	-3.6858***	(0.1252)
Alternative specific constants					
Non-participation	f_1	0.4954***	(0.0107)	0.2960***	(0.0266)
Full-time	f_4/f_3	0.6419***	(0.0061)	1.5241***	(0.0081)
Number of Observations		356,615		305,722	

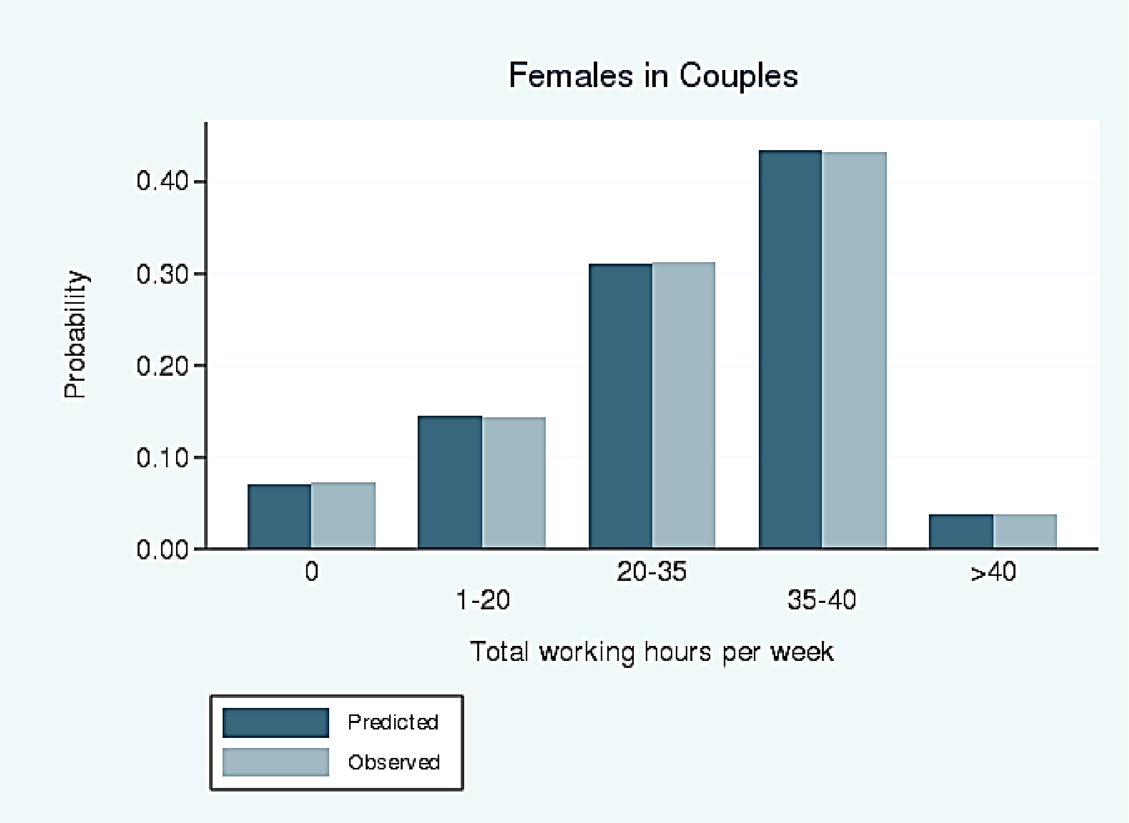
Note: *significant at 0.10 level, ** significant at 0.05 level, ***significant at 0.01 level

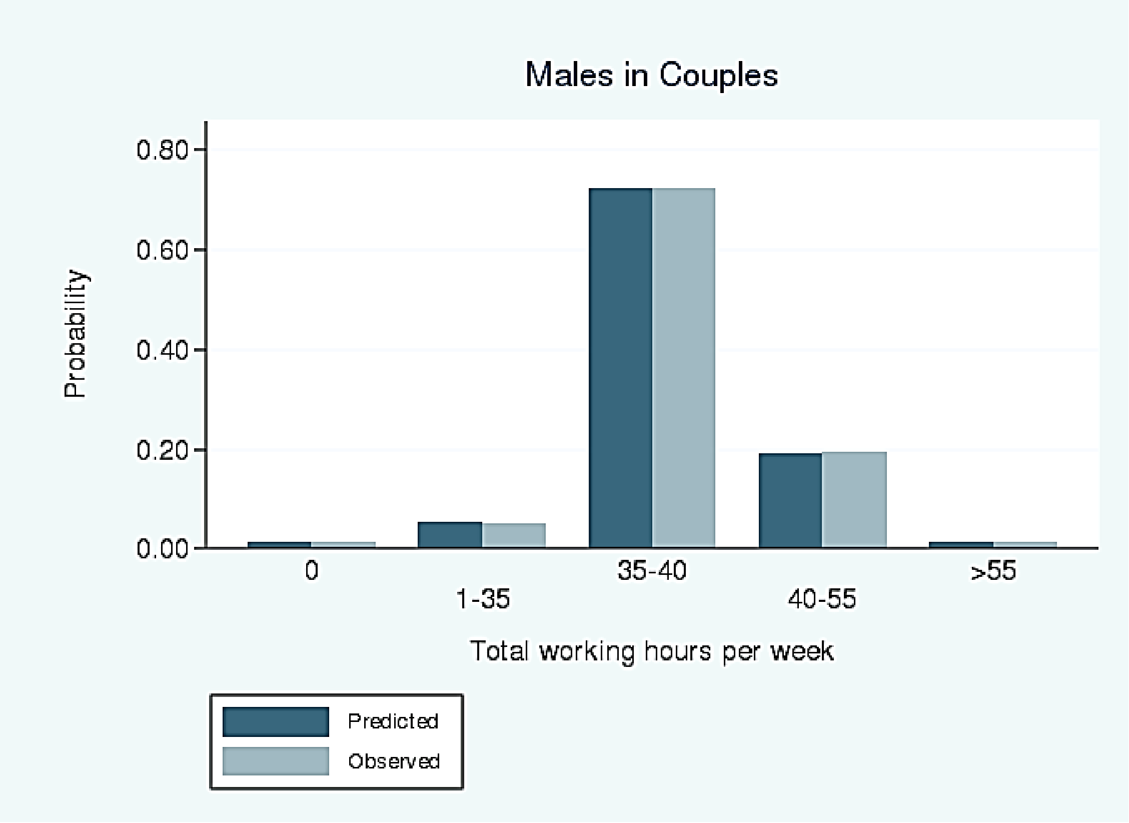
Figure A.1. Predicted and observed probabilities for working hours



Single Females

Single Males


Females in Couples
Probability
0.00
0.10
0.20
0.30
0.40
0
1-20
20-35
35-40
>40
Total working hours per week
Predicted
Observed


Males in Couples
Probability
0.00
0.20
0.40
0.60
0.80
0
1-35
35-40
40-55
>55
Total working hours per week
Predicted
Observed

Appendix B. Supplements to the NTR estimation

Summary Statistics

Table B.1. Pooled summary statistics 2000–2008 by gender and marital status. Mean and standard deviation in parentheses

	Single females	Single males	Females, couple	Males, couple
Total working hours	34.7 (8.08)	38.4 (7.15)	31.1 (8.8)	38.5 (6.2)
Labor income (2004 NOK)	314,665	390,184	280,052	443,678
Labor income (norm.)	1.00 (0.34)	1.24 (0.51)	0.89 (0.36)	1.41 (0.67)
Socioeconomic characteristics				
Wealth	4.24 (5.75)	4.04 (5.79)	6.00 (6.18)	5.47 (6.06)
Age	41.5 (10.5)	39.7 (10.3)	43.9 (9.6)	45.0 (9.5)
Married			0.84 (0.37)	0.83 (0.37)
Newborn	0.02 (0.16)	0.01 (0.09)	0.22 (0.48)	0.27 (0.53)
No. children under 6	0.06 (0.26)	0.01 (0.14)	0.41 (0.70)	0.47 (0.74)
No. children above 6	0.33 (0.65)	0.07 (0.32)	0.76 (0.96)	0.77 (0.97)
Non-western origin	0.03 (0.18)	0.04 (0.20)	0.04 (0.20)	0.04 (0.20)
Residence in Oslo	0.32 (0.47)	0.28 (0.45)	0.21 (0.41)	0.22 (0.41)
Densely populated area	0.87 (0.34)	0.82 (0.38)	0.78 (0.41)	0.81 (0.39)
Years of education	12.8 (2.60)	12.6 (2.56)	12.6 (2.60)	12.9 (2.66)
Field of education				
General	0.22 (0.41)	0.21 (0.41)	0.23 (0.42)	0.18 (0.39)
Human, art	0.06 (0.25)	0.04 (0.19)	0.05 (0.22)	0.03 (0.18)
Education	0.11 (0.32)	0.04 (0.20)	0.13 (0.34)	0.06 (0.24)
Social, law	0.04 (0.19)	0.03 (0.17)	0.03 (0.16)	0.03 (0.18)
Business, administration	0.18 (0.38)	0.11 (0.32)	0.17 (0.38)	0.12 (0.32)
Natural sciences, technology	0.07 (0.26)	0.41 (0.49)	0.06 (0.24)	0.42 (0.49)
Health	0.27 (0.44)	0.04 (0.20)	0.28 (0.45)	0.04 (0.20)
Primary industries	0.01 (0.08)	0.02 (0.13)	0.01 (0.07)	0.02 (0.14)
Service	0.02 (0.14)	0.06 (0.24)	0.02 (0.13)	0.06 (0.25)
Number of observations	1,325,331	1,330,061	3,014,522	2,699,336

Robustness checks

A main difficulty with the NTR approach is, as already discussed, to distinguish between the effect of the tax reform and the mean reversion effect that would be present also in absence of a tax reform. When only analyzing one period before and after the tax reform, it is hard to distinguish the two effects. We therefore rely on estimating the model over a period both with and without tax reform (2000–2008) to improve the coefficients of the mean reversion controls. A simple check of model performance is to compare the estimates of the control variables from the main model with estimates from a simplified ordinary least square regression over the period without tax changes (2000–2005). We find it reassuring that the coefficients for mean reversion and the other control variables are almost identical in the two regressions. Full regression outputs are available upon request.

In the following we discuss the effects on NTR estimates of some of the restrictions enforced by the main empirical strategy. First, in the main analysis we limit the sample to individuals with labor earnings above percentile 33 (about NOK250,000 in 2004) in the base year (the first year in each three-year difference) and working time above 30 working hours per week. The reasons for excluding observations at lower income and working hour levels is that the mean reversion problem is especially severe for individuals who initially have low income. This combined with the fact that the changes in marginal tax rates of the 2006 tax reform affected tax-payers at high income levels, makes the lowest income individuals redundant for identification. In Table B.2 we present estimation results for different sample restrictions. We find that the estimates are larger and more unstable with regards to the choice of mean reversion control when we do not have any sample restrictions.[38] However, results seem to be relatively insensitive to the exact choice of cut-off at percentile 33 and working hours ≥ 30. One exception is that working hours elasticties seem to decrease with the cut-off level for the income percentile between percentile 25 and 40. This might be due to average working hours responses of the reform depending on the hourly wage rate of the affected individuals. Keep in mind that the same cut-off rule is used for converting the structural model simulations into comparable net-of-tax elasticties, so this fact is not crucial for the validation exercise.

Table B.2. Estimates of net-of tax rate elasticities for alternative data restrictions

	Working hours		Earned income	
	Net-of-tax rate elasticity	Std error	Net-of-tax rate elasticity	Std error
Benchmark (≥30 working hours, ≥33 income percentile)	0.038	(0.0024)	0.055	(0.0022)
≥0 working hours	0.060	(0.0023)	0.052	(0.0022)
≥25 working hours	0.040	(0.0023)	0.054	(0.0022)
≥35 working hours	0.037	(0.0024)	0.059	(0.0023)
≤60 working hours	0.039	(0.0023)	0.054	(0.0022)
≥0 income percentile	0.089	(0.0024)	0.101	(0.0022)
≥25 income percentile	0.058	(0.0024)	0.052	(0.0022)
≥40 income percentile	0.019	(0.0024)	0.055	(0.0022)
≤99 income percentile	0.041	(0.0024)	0.047	(0.0023)
Number of observations	5,486,168	4,933,291	5,486,168	4,933,291

Note: All regressions include control variables for gender, wealth, age, age squared, married, number of children under and above the age of 6, newborn, residence in Oslo/ densely populated area, non-western origin, years of education, dummies for field of education, year dummies and third degree polynomial of working hours or labor income respectively. A sample restriction of ≥30 working hours and ≥33 income percentile in the base year is used as benchmark in order to avoid groups in which mean reversion is especially pronounced and therefore serve as a poor control group for the tax reform studied.

Next, we present robustness checks regarding the choice of time span. The three-year span has been proposed in the literature to allow some time for individuals to respond to tax changes. As already noted, this is an ad hoc choice (initiated by Feldstein, 1995), and in Table B.3 we present the

[38] Our interpretation of this result is that the problem of mean reversion is so severe for the low income individuals that including them in the regression gives us biased results of the tax effects.

results for alternative spans: one to four years. The elasticity estimates increase somewhat with the time span, in particular with respect to earnings. The likely reason is that wage earners respond to tax changes with some lag.

Table B.3. Robustness checks for time span assumption, net-of-tax rate elasticities for working hours and earned income

	Working hours		Earned income		
	Net-of-tax rate elasticity	Std error	Net-of-tax rate elasticity	Std error	Number of observations
Benchmark (three years)	0.038	(0.0023)	0.055	(0.0022)	2,648,201
One year	0.039	(0.0034)	0.025	(0.0021)	4,116,871
Two years	0.037	(0.0026)	0.040	(0.0021)	3,324,602
Four years	0.059	(0.0025)	0.058	(0.0025)	2,076,707

Note: All regressions include control variables for gender, wealth, age, age squared, married, number of children under and above the age of 6, newborn, residence in Oslo/ densely populated area, non-western origin, years of education, dummies for field of education, year dummies and third degree polynomial of working hours or labor income respectively.

As discussed in Section 5, the NTR estimates are obtained from an econometric specification without a representation of income effects. The main reason is that a collinearity problem materializes, as the instrument for the income effect is constructed in basically the same way as the instrument for the net-of-tax rate. However, in Table B.4 we report the results when explicitly accounting for income effects in the regressions for working hours, using the approach suggested by Blomquist and Selin (2010) to establish virtual income. We see that the income effect is small, and not necessarily negative as expected, and correspondingly, the effects of including income effects on the uncompensated and implied compensated net-of-tax elasticties are modest.

Table B.4. Net-of-tax rate elasticity estimates of specifications with and without income effects

	Uncompensated net-of-tax rate elastictity	Non-labor income elasticty	Implied compensated net-of-tax rate elastictity
Single females	0.0324***		0.0324***
	0.0577***	0.0068***	0.0512***
Single males	0.0227***		0.0227***
	0.0333***	0.0006	0.0327***
Females, couple	0.0514***		0.0514***
	0.0433***	-0.0191***	0.0615***
Males, couple	0.0160***		0.0160***
	0.0099**	-0.0035	0.0132***

Note: The implied compensated net-of-tax elasticity is estimated by the formula $\xi^C = \xi^U - \xi^R\left((1-\tau)q/R\right)$ where ξ^C, ξ^U and ξ^R refer to the compensated, uncompensated and non-labor income elasticity, respectively, see Blomquist and Selin (2010).