Essays on the simulation-based estimation of dynamic discrete choice models



Ben Waltmann

Balliol College

University of Oxford

Thesis submitted in fulfilment of the requirements for the degree of

Doctor of Philosophy in Economics

Trinity 2021

Introduction

A common complaint about structural work in applied microeconomics is that its methods are insufficiently transparent and its findings are not robust enough (see e.g. Angrist and Pischke, 2010). This critique seems especially on target in the context of complex dynamic models with large numbers of parameters, which are the subject of this thesis. In this literature, the important consequences of seemingly technical assumptions are often opaque even to the researchers themselves. Any two researchers may make different choices about how to specify and estimate a model, leading to different results.

For the most part, the target of this critique is the specification of the structural models themselves. Critics complain that models are likely to be misspecified; that they rely on unrealistic assumptions about agent knowledge and rationality; or that the model parameters are not identified given the data, even assuming that the model captures the true structure of the data generating process. These concerns have received a lot of attention in recent work: many structural models now feature deviations from full information or rationality (e.g. Abaluck and Adams-Prassl, 2021), new econometric techniques promise more transparency about the effects of identifying assumptions (e.g. Andrews, Gentzkow and Shapiro, 2017, 2020), and it is now much more common in applied work to see explicit (formal or informal) arguments about identification than it was in the 1990s.

A separate issue, which has received less attention, is a lack of transparency and insufficient robustness of parameter estimates stemming from estimation procedures. Even if a model is well-specified and (point-)identified, parameter estimates may be far away from the true parameters if estimators are biased or inefficient. Furthermore, the limits of computational precision can lead to estimation 'failures' even when estimators would be efficient with unlimited precision. These issues are especially prominent for dynamic discrete choice models, which are particularly challenging to estimate.

Achieving reasonable efficiency and robustness in the estimation of complex dynamic discrete choice models often requires substantial computational resources. This creates sharp trade-offs for applied researchers between computational costs, the accuracy of parameter estimates, and the complexity of models that can be estimated. Despite large advances in computing power, it is still common to estimate models relying on imprecise approximations, which can introduce economically meaningful bias into the parameter estimates. As a consequence,

the results of counterfactual policy experiments can also be misleading.

In this thesis, I propose new tools for estimating these models with a higher degree of accuracy at any given level of computational resources and model complexity. The thesis consists of this introduction, three separate research papers, and a conclusion. The first paper proposes a new technique for simulating the likelihood in finite-horizon dynamic discrete choice models with observed payoffs that makes better use of the information contained in observed payoffs to guide the simulation of choice probabilities. The second paper (with Jack Britton) proposes an extension to the Generalized Indirect Inference (GII) method of Bruins et al. (2018), which makes it practicable to estimate complex dynamic discrete choice models using GII. We illustrate our method by employing it to estimate a model of the career choices of young people in the UK, which we use to assess the long-run effects of a conditional cash transfer scheme. The third paper (also with Jack Britton) proposes three improvements to the approximate solution method of Keane and Wolpin (1994), which lead to much higher accuracy without increasing the required computation time. Unfortunately, the presentation as separate research papers means that a small amount of material is repeated across papers/chapters.

Acknowledgements. I would like to thank my advisers, James Duffy and Guido Ascari, for their invaluable guidance and encouragement in preparing this thesis; I hope Guido will forgive me that my academic interests have somewhat diverged from his over the course of the DPhil. I am further indebted to my examiners, Hamish Low and Aureo de Paula, whose commets have been extremely valuable. I would also like to thank Jack Britton, my co-author and now line manager at the Institute for Fiscal Studies. Jack first introduced me to the topics explored in this thesis in 2017 and has been an invaluable source of advice and guidance since. I am grateful to Balliol College and the Economic and Social Research Council for funding in the form of a Doctoral Studentship, and to the Institute for Fiscal Studies for granting me leave to work on this thesis in its final stages.

The most important thank you, as ever, belongs to my family for their love and support. I am also deeply grateful to Jan David Bakker, Friedemann Bieber, Sofia Birattari, Hélène Colinet, Joshua Folkard, Thitapa (Ploy) Haritaworn, Mimi Lu, Catherine Papastavrou-Brooks, Markus Parlasca, Margherita Philipp, Katia Schwichtenberg, Kardin Sømme, Isabel Stockton and Xiaowei Xu. I could not have done this without you.

References

Abaluck, Jason, and Abi Adams-Prassl. 2021. "What do Consumers Consider Before They Choose? Identification from Asymmetric Demand Responses*." *The Quarterly Journal of Economics*, 136(3): 1611–1663.

Andrews, Isaiah, Matthew Gentzkow, and Jesse M Shapiro. 2017. "Measuring the sensitivity of parameter estimates to estimation moments." *The Quarterly Journal of Economics*, 132(4): 1553–1592.

Andrews, Isaiah, Matthew Gentzkow, and Jesse M Shapiro. 2020. "On the informativeness of descriptive statistics for structural estimates." *Econometrica*, 88(6): 2231–2258.

Angrist, Joshua D, and Jörn-Steffen Pischke. 2010. "The credibility revolution in empirical economics: How better research design is taking the con out of econometrics." *Journal of economic perspectives*, 24(2): 3–30.

Bruins, Marianne, James A Duffy, Michael P Keane, and Anthony A Smith Jr. 2018. "Generalized indirect inference for discrete choice models." *Journal of Econometrics*, 205(1): 177–203.

Keane, Michael P, and Kenneth I Wolpin. 1994. "The solution and estimation of discrete choice dynamic programming models by simulation and interpolation: Monte Carlo evidence." *The Review of Economics and Statistics*, 648–672.

Summary of Contents

I. **Efficient likelihood simulation in discrete choice models with observed payoffs**

In many discrete choice contexts, some information about the agent's payoff is observed along with her choice. Simulated Maximum Likelihood estimation then requires simulation of the joint density of the agent's choices and observed payoffs. High simulation accuracy is crucial to minimize bias in parameter estimates, yet in many cases practically infeasible using existing simulation methods. This paper proposes a new simulator that makes better use of the relationship between choices and payoffs implied by the model. The new simulator strictly dominates existing methods in terms of accuracy and computation time. When a canonical model of career choice is estimated using the new simulator, parameter estimates are more accurate and bias in a counterfactual simulation is lower than when existing methods are used. The new method can accommodate both unobserved state variables and classification error.

II. **Generalized indirect inference in dynamic structural models: an application to the long-run effect of a conditional cash transfer scheme** (with Jack Britton)

The Generalized Indirect Inference (GII) method of Bruins et al. (2018) substantially eases the computational burden of estimating discrete choice models using Indirect Inference. However, it has so far been impossible to apply the GII method to an important class of structural dynamic discrete choice models that feature both unobserved heterogeneity and payoffs associated with different options that can depend on a long history of past choices. This paper shows how the GII method can be adapted to make the estimation of such models possible. We compare our method to Simulated Maximum Likelihood and conventional GII in a structural dynamic discrete choice setting where all three methods are feasible. Our new method roughly matches the other two in terms of both estimation accuracy and performance in a simple policy experiment. We then use the method to estimate a structural model of a conditional cash transfer scheme in the UK, which could not be estimated using either of the other methods. We find that the programme, which paid young people up to £3,600 to stay in education or training beyond age 16, increased attendance among students eligible for the full grant by up to 2.2 percentage points. While the policy only had a long-run impact for a small minority of students, wages for that group were 9 percent higher at age 26 than in a counterfactual scenario without the grant.

III. **Revisiting the solution of dynamic discrete choice models: time to bring back Keane and Wolpin (1994)?** (with Jack Britton)

The 'curse of dimensionality' is a common problem in the estimation of dynamic models: as models get more complex, the computational cost of solving these models rises exponentially. Keane and Wolpin (1994) proposed a method for addressing this problem in finite-horizon dynamic discrete choice models by evaluating only a subset of state space points by Monte Carlo integration and interpolating the value of the remainder. This method was widely used in the late 1990s and 2000s but has rarely been used since, as it was found to be unreliable in some settings. In this paper, we develop an improved version of their method that relies on three amendments: systematic sampling, data-guided selection of state space points for Monte Carlo integration, and dispensing with polynomial interpolation when a multicollinearity problem is detected. With these improvements, the Keane and Wolpin (1994) method achieves excellent approximation performance even in a model with a large state space and substantial *ex ante* heterogeneity.

I. Efficient likelihood simulation in discrete choice models with observed payoffs*

BEN WALTMANN
Oxford and IFS

Abstract

In many discrete choice contexts, some information about the agent's payoff is observed along with her choice. Simulated Maximum Likelihood estimation then requires simulation of the joint density of the agent's choices and observed payoffs. High simulation accuracy is crucial to minimize bias in parameter estimates, yet in many cases practically infeasible using existing simulation methods. This paper proposes a new simulator that makes better use of the relationship between choices and payoffs implied by the model. The new simulator strictly dominates existing methods in terms of accuracy and computation time. When a canonical model of career choice is estimated using the new simulator, parameter estimates are more accurate and bias in a counterfactual simulation is lower than when existing methods are used. The new method can accommodate both unobserved state variables and classification error.

*I am grateful Jan Bakker, Jack Britton, James Duffy, David Goll, Mike Keane, Hamish Low, Markus Parlasca, Robert Sauer, Michela Tincani, and seminar participants at UCL and Oxford for useful comments and suggestions. The use of the University of Oxford Advanced Research Computing (ARC) facility is gratefully acknowledged (DOI code: http://dx.doi.org/10.5281/zenodo.22558). Email: ben.waltmann@ifs.org.uk.

1 Introduction

This paper proposes a new method for simulating the likelihood in a popular class of discrete choice models in which for some of the available options, an agent's flow payoff is observed (with error) when that option is chosen. Models within this class include forward-looking dynamic career choice models in the tradition of Keane and Wolpin (1994). In these models, agents choose between (different varieties of) schooling, market work, and non-market work. If and only if an agent chooses market work, their current payoff is observed in the form of their wage, usually subject to measurement error. Examples of papers in this tradition are Keane and Wolpin (1997), Eckstein and Wolpin (1999), Keane and Wolpin (2001), Stinebrickner (2001), Sauer (2004), Todd and Wolpin (2006), Cohen-Goldner and Eckstein (2008), and Sauer (2015).

There is typically no analytical form available for the likelihood in models of this class. Hence Maximum Likelihood estimation requires the use of numerical integration methods; typically, Monte Carlo integration is used. Many different likelihood simulators are consistent but vary in their efficiency. This paper considers two commonly used simulators from the literature and proposes a new simulator that strictly dominates both in terms of computation time and efficiency.

The proposed new simulator uses information from the observed wage to guide Monte Carlo integration. More samples are drawn in regions of the sample space that are more likely to have generated the data. This improves the efficiency of the simulator relative to the two other simulators found in the literature (hereafter called the 'Simple Simulator' and the 'Wolpin Simulator'), which do not explicitly take into account the relationship between payoffs and choices at the sampling stage. The standard version of the Simple Simulator is used in Sauer (1998) and Cohen-Goldner and Eckstein (2008). An extended version developed by Keane and Wolpin (2001) and Keane and Sauer (2010) allows for classification error and unobserved state variables. This version is employed in Sauer (2004) and Sauer (2015). The 'Wolpin Simulator' is used in Eckstein and Wolpin (1999) and Todd and Wolpin (2006).[1]

Monte Carlo evidence shows that for a fixed number of draws, the new simulator yields more accurate density evaluations than the two alternatives for all variations of the model tested. The accuracy difference compared to the Simple Simulator is particularly large when

[1] Keane and Wolpin (1994) and Keane and Wolpin (1997) contain no explicit account of the simulation procedure used.

the measurement error in observed payoffs is relatively small, whereas the difference to the Wolpin Simulator is large when the measurement error in observed payoffs is relatively large. The accuracy improvement from using the new simulator increases with higher correlation of structural disturbances across alternatives.

This improved accuracy in likelihood simulation translates into more precisely estimated parameters in Simulated Maximum Likelihood estimation. In particular, I show that when used to estimate a reasonably calibrated version of the canonical career choice model of Keane and Wolpin (1994), the new method roughly halves the root mean squared error of estimated parameters compared to existing methods. More accurate parameter estimates in turn enable more accurate counterfactual policy simulations. I show that a researcher studying the effect of a tuition subsidy on choice shares using the Keane and Wolpin (1994) model would typically have found an effect within 15 percent of the true magnitude if they estimated the model using the new simulator, but would typically have been more than 30 percent away from the true magnitude if they had used existing methods from the literature.

The same qualitative results obtain with unobserved state variables; indeed the differences between the simulators are larger and grow with the number of periods T for which data is observed. When T is large, a Sequential Monte Carlo approach, which can be combined with any of the simulators, leads to large improvements in simulation accuracy of all simulators even if the number of simulation draws is reduced to account for the added computational cost (the relative performance of the simulators is unaffected).

When there is classification error in the data, the new simulator becomes more costly to compute and loses efficiency. Hence the Simple Simulator will be preferable for some combinations of classification error, measurement error in wages, and correlation between the structural disturbances. However, for most plausible values of all three, the new simulator still yields more accurate results even with computation time held fixed. Furthermore, the new simulator is more robust than the Simple Simulator, as it provides acceptable accuracy across all possible scenarios.

The rest of the paper is structured as follows. Section 2 outlines the general model that guides the subsequent discussion. Section 3 derives the standard versions of the three simulators and compares their performance using Monte Carlos simulations. The new simulator is extended to allow for unobserved state variables in Section 4 and classification error in models with and without unobserved state variables in Section 5. Section 6 shows that the higher

simulation accuracy of the new simulator translates into more precisely estimated parameters and reduced bias in counterfactual simulations. Section 7 concludes.

2 General model

This section introduces a general model that structures the subsequent discussion. All random disturbances in the model are jointly independent across agents as well as across periods for the same agent.[2] In each period t, an agent i chooses one among a menu of N mutually exclusive options. She chooses the option with the highest (expected) payoff. The payoff of an option n is

$$V_{nit} = \Phi_n(E_{nit}; \boldsymbol{s}_{it}) + \Psi_n(\boldsymbol{s}_{it}). \tag{1}$$

Φ_n is a known function that is invertible with respect to its first argument. E_{nit} is the nth element of a vector of serially uncorrelated structural shocks $\boldsymbol{E}_{it}$ with variance $\boldsymbol{\Sigma}$ that is observed by the agent in the current period but not by the econometrician. Ψ_n is a known function of an exogenous vector of observed state variables $\boldsymbol{s}_{it}$ that is treated as deterministic. Φ_n can be interpreted as the agent's flow payoff, and Ψ_n as the continuation value of her choice.

In each period and for each individual, the econometrician observes the vector of state variables $\boldsymbol{s}_{it}$, the realization $\boldsymbol{d}_{it}$ of the true choice vector $\boldsymbol{D}_{it}$ and the observed flow payoff w^*_{it}, which is the realization of the scalar

$$W^*_{it} = \boldsymbol{D}_{it}\boldsymbol{\Omega}\boldsymbol{V}^*_{it}. \tag{2}$$

$\boldsymbol{D}_{it} = (D_{1it}, D_{2it}, \ldots, D_{Nit})'$ is a vector of indicator variables: $D_{nit} = 1$ if the agent chooses option n and zero otherwise.[3] $\boldsymbol{\Omega}$ is a diagonal matrix with the elements $\boldsymbol{\omega} = (\omega_1, \omega_2, \ldots, \omega_N)$ along its main diagonal, where $\omega_n = 1$ if the payoff for choice n is observed (with error) and $\omega_n = 0$ otherwise. $\boldsymbol{V}^*_{it}$ is the vector whose nth element is

$$V^*_{nit} = \Phi_n(E_{nit} + H_{nit}; \boldsymbol{s}_{it}) \tag{3}$$

[2] This assumption is common in the literature (Keane and Wolpin, 1994, 1997, 2001; Eckstein and Wolpin, 1999; Sauer, 2004, 2015; Todd and Wolpin, 2006; Cohen-Goldner and Eckstein, 2008). It is relaxed in section 4.

[3] The assumption that choices are observed without error is also common in the literature (Keane and Wolpin, 1997; Eckstein and Wolpin, 1999; Todd and Wolpin, 2006; Cohen-Goldner and Eckstein, 2008). Indeed, I am aware of no paper in this literature that assumes classification error when this is not crucial for the feasibility of the estimation procedure. Nevertheless, the assumption is relaxed in section 5.

where H_{nit} (the measurement error) is the nth element of a random vector $\boldsymbol{H}_{it}$ that is uncorrelated with the structural shock. $\boldsymbol{V}^*_{it}$ can be interpreted as the vector of latent observed (flow) payoffs. Let $V^*_{it}(\boldsymbol{d}_{it})$ refer to the element of $\boldsymbol{V}^*_{it}$ corresponding to the chosen option, i.e.

$$V^*_{it}(\boldsymbol{d}_{it}) = \boldsymbol{d}'_{it}\boldsymbol{V}^*_{it}. \tag{4}$$

As is standard in the literature (Keane and Wolpin, 1994, 1997, 2001; Eckstein and Wolpin, 1999; Sauer, 2004, 2015; Todd and Wolpin, 2006; Cohen-Goldner and Eckstein, 2008), assume that each $\boldsymbol{E}_{it}$ is drawn from the multivariate normal distribution $N(0, \boldsymbol{\Sigma})$ and that the elements of $\boldsymbol{H}_{it}$ are independent and drawn from the univariate normal distribution $N(0, \sigma^2_\eta)$.

This form is general enough to accommodate many models in the discrete choice literature.[4] A canonical model of this type is the Keane and Wolpin (1994) model of occupational choice. In that model, $N = 4$, $\omega = (1, 1, 0, 0)$ and w^*_{it} is the observed wage. The flow payoff Φ_n takes the form

$$\Phi_n(E_{nit}; \boldsymbol{s}_{it}) = \begin{cases} \exp(\Upsilon_n(\boldsymbol{s}_{it}) + E_{nit}) & \text{for } n = 1, 2 \\ \Upsilon_n(\boldsymbol{s}_{it}) + E_{nit} & \text{for } n = 3, 4. \end{cases} \tag{5}$$

where Υ_n is a function of the observed state $\boldsymbol{s}_{it}$. $\Psi_n(\boldsymbol{s}_{it})$ is the expected value of future flow payoffs.[5]

3 Simulating the likelihood

A common method of estimating models of this class is Simulated Maximum Likelihood. Since there are no unobserved state variables and the structural shocks are assumed to be serially uncorrelated, the likelihood contributions of an individual's choices and observed payoffs can be separated between periods. The per-period density of the data is $f_{\boldsymbol{D},W^*}(\boldsymbol{d}_{it}, w^*_{it})$.[6]

As there is in general no analytical form available for this expression, it usually has to be simulated. How exactly $f_{\boldsymbol{D},W^*}(\boldsymbol{d}_{it}, w^*_{it})$ is simulated can make a large difference to simulation accuracy and thus the efficiency of the Simulate Maximum Likelihood estimator, yet it is rarely explicitly discussed in the literature. In this section, I start by laying out how two com-

[4] All of the arguments in this paper also apply to an even more general model where $V_{nit} = \Gamma_n(E_{nit}, \boldsymbol{s}_{it})$, and Γ_n is any known function. (1) is given the more specific form to retain the interpretation of w^*_{it} as an observed (flow) payoff.

[5] A more detailed exposition of the Keane and Wolpin (1994) model is given in section 6.1.

[6] Subscripts of subscripts are dropped throughout when this does not causes ambiguity. The state s_{it} is treated as deterministic, so I do not explicitly condition on it.

mon simulators from the literature are constructed, which I call the 'Simple Simulator' and the 'Wolpin Simulator'. I then propose a third simulator (the new simulator), which achieves greater efficiency. Finally, I provide Monte Carlo evidence that both simulators from the literature perform reasonably well in particular circumstances, but the new simulator strictly dominates both in terms of computation time and efficiency.[7]

Like all simple frequency simulators, the simulators of $f_{\boldsymbol{D},W^*}(\boldsymbol{d}_{it}, w^*_{it})$ discussed in the main part of this paper would all be *non-smooth* functions of any parameter vector. There are two strategies available for obtaining smooth versions of these simulators. First, the familiar kernel smoothing approach of McFadden (1989) can be used, which introduces bias. A second strategy, pursued in Appendix A, is to construct an unbiased and consistent version of these simulators using a variation of the GHK algorithm (Keane, 1994).

3.1 Simple Simulator

The density $f_{\boldsymbol{D},W^*}(\boldsymbol{d}_{it}, w^*_{it})$ can be written as

$$f_{\boldsymbol{D},W^*}(\boldsymbol{d}_{it}, w^*_{it}) = \int f_{\boldsymbol{D},W^*|\boldsymbol{E}}(d_{it}, w^*_{it}|\boldsymbol{\epsilon}_{it})\phi_{\boldsymbol{E}}(\boldsymbol{\epsilon}_{it})d\boldsymbol{\epsilon}_{it} \tag{6}$$

where $\boldsymbol{E}_{it}$ is the vector of structural shocks whose particular realization is $\boldsymbol{\epsilon}_{it}$.[8] An equivalent expression is

$$f_{\boldsymbol{D},W^*}(\boldsymbol{d}_{it}, w^*_{it}) = \int f_{W^*|\boldsymbol{E}}(w^*_{it}|\boldsymbol{\epsilon}_{it})P(\boldsymbol{D}_{it} = \boldsymbol{d}_{it}|\boldsymbol{E}_{it} = \boldsymbol{\epsilon}_{it})\phi_{\boldsymbol{E}}(\boldsymbol{\epsilon}_{it})d\boldsymbol{\epsilon}_{it}. \tag{7}$$

As there is no information about choices in the observed wage over and above the vector of shocks and the state vector, there is no need to condition on w^*_{it} in the expression for the choice probability in (7).[9] A consistent simulator based on (7) is

$$\hat{f}^S_{\boldsymbol{D},W^*}(\boldsymbol{d}_{it}, w^*_{it}) = \frac{1}{R}\sum_{r=1}^{R} f_{W^*|\boldsymbol{E}}(w^*_{it}|\boldsymbol{\epsilon}^r_{it})P(\boldsymbol{D}_{it} = \boldsymbol{d}_{it}|\boldsymbol{E}_{it} = \boldsymbol{\epsilon}^r_{it}) \tag{8}$$

[7] Note that if the agent chooses an option where no payoff is observed (i.e. $\boldsymbol{d}'_{it}\boldsymbol{\omega} = 0$), w^*_{it} will be zero by construction. Hence w^*_{it} contains no information and the likelihood contribution is $P(\boldsymbol{D}_{it} = \boldsymbol{d}_{it})$. This paper is only concerned with simulation in cases where a payoff is observed (i.e. $\boldsymbol{d}'_{it}\boldsymbol{\omega} \neq 0$).

[8] Equivalently, one could condition on $\boldsymbol{V}_{it}$. This is essentially a matter of notation, since given $\boldsymbol{s}_{it}$, there is a one-to-one correspondence between $\boldsymbol{V}_{it}$ and $\boldsymbol{E}_{it}$.

[9] Alternatively, one could reason that there is no information in $\boldsymbol{d}_{it}$ about w^*_{it} over and above what is contained in $\boldsymbol{\epsilon}_{it}$.

where each $\boldsymbol{\epsilon}_{it}^r$ is drawn from its unconditional distribution. Then

$$P(\boldsymbol{D}_{it} = \boldsymbol{d}_{it} | \boldsymbol{E}_{it} = \boldsymbol{\epsilon}_{it}^r) = I(v_{cit}^r > v_{jit}^r \forall\, j \neq c) \tag{9}$$

where c is the index of the chosen option (i.e. $d_{cit} = 1$) and

$$v_{nit}^r = \Phi_n(\boldsymbol{\epsilon}_{nit}^r; \boldsymbol{s}_{it}) + \Psi_n(\boldsymbol{s}_{it})$$

for all n. $f_{W^*|\boldsymbol{E}}(w_{it}^* | \boldsymbol{\epsilon}_{it}^r)$ is the density of the observed payoff conditional on the simulated payoffs. Under the assumptions above, $f_{W^*|\boldsymbol{E}}$ will be a transformation of a Normal pdf with mean zero and variance σ_η^2, transformed by Φ_n.[10]

3.2 Wolpin Simulator

Let $\tilde{\boldsymbol{E}}_{it}$ be the vector that contains all elements of $\boldsymbol{E}_{it}$ except the element corresponding to the agent's chosen option, E_{cit}, which has been replaced by the measurement error H_{cit}. Then

$$f_{\boldsymbol{D},W^*}(\boldsymbol{d}_{it}, w_{it}^*) = \int f_{\boldsymbol{D},W^*|\tilde{\boldsymbol{E}}}(\boldsymbol{d}_{it}, w_{it}^* | \tilde{\boldsymbol{\epsilon}}_{it}) \phi_{\tilde{\boldsymbol{E}}}(\tilde{\boldsymbol{\epsilon}}_{it}) d\tilde{\boldsymbol{\epsilon}}_{it}. \tag{10}$$

An equivalent expression is

$$f_{\boldsymbol{D},W^*}(\boldsymbol{d}_{it}, w_{it}^*) = \int f_{W^*|\tilde{\boldsymbol{E}}}(w_{it}^* | \tilde{\boldsymbol{\epsilon}}_{it}) P(\boldsymbol{D}_{it} = \boldsymbol{d}_{it} | \tilde{\boldsymbol{E}}_{it} = \tilde{\boldsymbol{\epsilon}}_{it}, W_{it}^* = w_{it}^*) \phi_{\tilde{\boldsymbol{E}}}(\tilde{\boldsymbol{\epsilon}}_{it}) d\tilde{\boldsymbol{\epsilon}}_{it}. \tag{11}$$

Another consistent simulator is therefore

$$\hat{f}_{\boldsymbol{D},W^*}^W(d_{it}, w_{it}^*) = \frac{1}{R} \sum_{r=1}^{R} f_{W^*|\tilde{\boldsymbol{E}}}(w_{it}^* | \tilde{\boldsymbol{\epsilon}}_{it}) P(\boldsymbol{D}_{it} = \boldsymbol{d}_{it} | \tilde{\boldsymbol{E}}_{it} = \tilde{\boldsymbol{\epsilon}}_{it}^r, W_{it}^* = w_{it}^*) \tag{12}$$

where each $\tilde{\boldsymbol{\epsilon}}_{it}^r$ is drawn from its unconditional distribution. There is a unique ϵ_{cit}^r that is consistent with a given $\tilde{\boldsymbol{\epsilon}}_{it}^r$ and w_{it}^*, namely

$$\epsilon_{cit}^r = \Phi_c^{-1}(w_{it}^*; \boldsymbol{s}_{it}) - \eta_{cit}^r. \tag{13}$$

[10] Note that this density will be degenerate when there is no measurement error. Then $f_{W^*|\boldsymbol{E}}(w_{it}^* | \boldsymbol{\epsilon}_{it}^r) = 0$ *almost surely* and the simulator will be inconsistent.

Hence one can calculate $\boldsymbol{v}_{it}^r$ and thus

$$P(\boldsymbol{D}_{it} = \boldsymbol{d}_{it} | \tilde{\boldsymbol{E}}_{it} = \tilde{\boldsymbol{\epsilon}}_{it}^r, W_{it}^* = w_{it}^*) = I(v_{cit}^r > v_{jit}^r \forall\, j \neq c). \tag{14}$$

$f_{W^*|\tilde{E}}(w_{it}^* | \tilde{\boldsymbol{\epsilon}}_{it})$ is now a transformation of the conditional density of the structural shock corresponding to the chosen option ϵ_{cit}. As the structural shocks are allowed to be correlated across options, the appropriate density potentially depends on all elements of $\tilde{\boldsymbol{\epsilon}}_{it}^r$.

3.3 New simulator

As a first, preliminary step, I show that the joint distribution of $\boldsymbol{D}_{it}$ and W_{it}^* is the same as the joint density of $\boldsymbol{D}_{it}$ and $V_{it}^*(\boldsymbol{d}_{it})$. Recall that $V_{it}^*(\boldsymbol{d}_{it})$ is the element of $\boldsymbol{V}_{it}^*$ corresponding to a given choice $\boldsymbol{d}_{it}$, as defined by (4). Thus $V_{it}^*(\boldsymbol{d}_{it})$ does not inherit any randomness related to the agent's choice. In contrast, W_{it}^* as defined by (2) is a function of $\boldsymbol{V}_{it}^*$ and $\boldsymbol{D}_{it}$, *both* of which are random. W_{it}^* will in general have a different *marginal* distribution from each element of $\boldsymbol{V}_{it}^*$. However, provided that $\boldsymbol{d}_{it}'\boldsymbol{\omega} \neq 0$, it follows from the definitions of $V_{it}^*(\boldsymbol{d}_{it})$ and W_{it}^* that

$$f_{W^*|\boldsymbol{D}}(w_{it}^* | \boldsymbol{d}_{it}) = f_{V^*(\boldsymbol{d})|\boldsymbol{D}}(w_{it}^* | \boldsymbol{d}_{it}). \tag{15}$$

In words, $V_{it}^*(\boldsymbol{d}_{it})$ and W_{it}^* will have the same distribution *conditional* on any given $\boldsymbol{d}_{it}$. Multiplying both sides of (15) by $P(\boldsymbol{D}_{it} = \boldsymbol{d}_{it})$ yields the required result:

$$f_{\boldsymbol{D},W^*}(\boldsymbol{d}_{it}, w_{it}^*) = f_{\boldsymbol{D},V^*(\boldsymbol{d})}(\boldsymbol{d}_{it}, w_{it}^*). \tag{16}$$

Hence the joint density of $\boldsymbol{D}_{it}$ and the $V_{it}^*(\boldsymbol{d}_{it})$ is identical to the joint density of $\boldsymbol{D}_{it}$ and W_{it}^*. This justifies working with $V_{it}^*(\boldsymbol{d}_{it})$ rather than the more complex W_{it}^* in what follows.

Second, as $\phi_{E|V^*(\boldsymbol{d})}(\boldsymbol{\epsilon}_{it} | w_{it}^*)$ is a positive density function on $\mathbb{R}^N$, (7) can be written as

$$f_{\boldsymbol{D},V^*(\boldsymbol{d})}(\boldsymbol{d}_{it}, w_{it}^*) = \int \frac{f_{V^*(\boldsymbol{d})|\boldsymbol{E}}(w_{it}^* | \boldsymbol{\epsilon}_{it}) P(\boldsymbol{D}_{it} = \boldsymbol{d}_{it} | \boldsymbol{E}_{it} = \boldsymbol{\epsilon}_{it}) \phi_{\boldsymbol{E}}(\boldsymbol{\epsilon}_{it})}{\phi_{E|V^*(\boldsymbol{d})}(\boldsymbol{\epsilon}_{it} | w_{it}^*)} \phi_{E|V^*(\boldsymbol{d})}(\boldsymbol{\epsilon}_{it} | w_{it}^*) d\boldsymbol{\epsilon}_{it} \tag{17}$$

which simplifies to

$$f_{\boldsymbol{D},V^*(\boldsymbol{d})}(\boldsymbol{d}_{it}, w_{it}^*) = f_{V^*(\boldsymbol{d})}(w_{it}^*) \int P(\boldsymbol{D}_{it} = \boldsymbol{d}_{it} | \boldsymbol{E}_{it} = \boldsymbol{\epsilon}_{it}) \phi_{E|V^*(\boldsymbol{d})}(\boldsymbol{\epsilon}_{it} | w_{it}^*) d\boldsymbol{\epsilon}_{it}. \tag{18}$$

Another consistent simulation estimator is therefore

$$\hat{f}^N_{\boldsymbol{D},V^*(\boldsymbol{d})}(\boldsymbol{d}_{it}, w^*_{it}) = \frac{f_{V^*(\boldsymbol{d})}(w^*_{it})}{R} \sum_{r=1}^{R} P(\boldsymbol{D}_{it} = \boldsymbol{d}_{it} | \boldsymbol{E}_{it} = \boldsymbol{\epsilon}^r_{it}) \tag{19}$$

where each $\boldsymbol{\epsilon}^r_{it}$ is drawn from the distribution of $\boldsymbol{E}_{it}$ *conditional on* $V^*_{it}(\boldsymbol{d}_{it}) = w^*_{it}$, and $f_{V^*(\boldsymbol{d})}(w^*_{it})$ is simply the unconditional density of $V^*_{it}(\boldsymbol{d}_{it})$ evaluated at w^*_{it}. (19) can be thought of as an *importance sampling* version of (8).

$$P(\boldsymbol{D}_{it} = \boldsymbol{d}_{it} | \boldsymbol{E}_{it} = \boldsymbol{\epsilon}^r_{it}) = I(v^r_{cit} > v^r_{jit} \forall\, j \neq c) \tag{20}$$

where c is the index of the chosen option, is evaluated just as for the Simple Simulator. The challenge is to draw $\boldsymbol{\epsilon}^r_{it}$ from the distribution of $\boldsymbol{E}_{it}$ conditional on $V^*_{it}(\boldsymbol{d}_{it}) = w^*_{it}$.[11]

Define

$$E^*_{cit} = E_{cit} + H_{cit} = \Phi_c^{-1}(V^*_{it}(\boldsymbol{d}_{it}); \boldsymbol{s}_{it}). \tag{21}$$

where c is the index of the chosen option. The unconditional distribution of E^*_{cit} is $N(0, \sigma_c^2 + \sigma_\eta^2)$, where σ_c^2 is the variance of the structural shock attached to the chosen option (i.e. the cth entry along the main diagonal of $\boldsymbol{\Sigma}$) and σ_η^2 is the variance of the measurement error. Furthermore,

$$\phi_{\boldsymbol{E}|V^*(\boldsymbol{d})}(\boldsymbol{\epsilon}_{it} | v^*_{cit}) = \phi_{\boldsymbol{E}|E^*_c}(\boldsymbol{\epsilon}_{it} | \epsilon^*_{cit}) \tag{22}$$

because ϵ^*_{cit} and v^*_{cit} contain the same information about $\boldsymbol{\epsilon}_{it}$. Since all the information about $\boldsymbol{\epsilon}_{it}$ in ϵ^*_{cit} is also contained in ϵ_{cit}, it is possible to draw from $\phi_{\boldsymbol{E}|V^*(\boldsymbol{d})}(\boldsymbol{\epsilon}_{it} | v^*_{cit})$ using a sequential procedure: First, draw an ϵ^r_{cit} from $\phi_{E_c|E^*_c}(\epsilon_{cit} | \epsilon^*_{cit})$. Second, draw the rest of $\boldsymbol{\epsilon}^r_{it}$ from $\phi_{\boldsymbol{E}|E_c}(\boldsymbol{\epsilon}_{it} | \epsilon_{cit})$.

An analytic form for $\phi_{E_c|E^*_c}(\epsilon_{cit} | \epsilon^*_{cit})$ can be obtained using Bayes' theorem

$$\phi_{E_c|E^*_c}(\epsilon_{cit} | \epsilon^*_{cit}) = \frac{\phi_{E_c}(\epsilon_{cit}) \phi_{E^*_c|E_c}(\epsilon^*_{cit} | \epsilon_{cit})}{\phi_{E^*_c}(\epsilon^*_{cit})}. \tag{23}$$

$\phi_{E_c}(\epsilon_{cit})$ is the pdf of the Normal distribution $N(0, \sigma_c^2)$. $\phi_{E^*_c|E}(\epsilon^*_{cit} | \epsilon_{cit})$ is the pdf of the Normal

[11] An alternative derivation of the new simulator relies on the fact that by definition,

$$f_{\boldsymbol{D},V^*(\boldsymbol{d})}(\boldsymbol{d}_{it}, w^*_{it}) = f_{V^*(\boldsymbol{d})}(w^*_{it}) P(\boldsymbol{D}_{it} = \boldsymbol{d}_{it} | V^*_{it}(\boldsymbol{d}_{it}) = w^*_{it})$$

and therefore

$$f_{\boldsymbol{D},V^*(\boldsymbol{d})}(\boldsymbol{d}_{it}, w^*_{it}) = f_{V^*(\boldsymbol{d})}(w^*_{it}) \int P(\boldsymbol{D}_{it} = \boldsymbol{d}_{it} | \boldsymbol{E}_{it} = \boldsymbol{\epsilon}_{it}) \phi_{E|V^*(\boldsymbol{d})}(\boldsymbol{\epsilon}_{it} | w^*_{it}) d\boldsymbol{\epsilon}_{it}.$$

The conditioning of the choice probability on $V^*_{it}(\boldsymbol{d}_{it})$ can be dropped because it adds no information regarding the choice probability over $\boldsymbol{E}_{it}$.

distribution $N(\epsilon_{cit}, \sigma_\eta^2)$. $\phi_{E_c^*}(\epsilon_{cit}^*)$ is the pdf of the Normal distribution $N(0, \sigma_c^2 + \sigma_\eta^2)$. It follows that $\phi_{E_c|E_c^*}(\epsilon_{cit}|\epsilon_{cit}^*)$ is the pdf of the Normal distribution[12]

$$N\left(\frac{\sigma_c^2}{\sigma_c^2+\sigma_\eta^2}\epsilon_{cit}^*, \frac{\sigma_c^2\sigma_\eta^2}{\sigma_c^2+\sigma_\eta^2}\right). \tag{24}$$

With this at hand, drawing an ϵ_{cit}^r from $\phi_{E_c|E_c^*}(\epsilon_{cit}|\epsilon_{cit}^*)$ is straightforward. The rest of $\boldsymbol{\epsilon}_{it}^r$ can be drawn as follows. First, re-order the covariance matrix $\boldsymbol{\Sigma}$ so that σ_c^2 is the first entry along the main diagonal; call that matrix $\boldsymbol{\Sigma}^c$. Second, construct a vector $\boldsymbol{n}_{it} = (n_{1it}, n_{2it}, \ldots, n_{Nit})$, where

$$n_{1it} = \epsilon_{cit}^r/\sigma_c. \tag{25}$$

and $(n_{2it}, \ldots, n_{Nit})$ are draws from the standard Normal distribution. Third, compute

$$\boldsymbol{\epsilon}_{it}^{rc} = \boldsymbol{\Gamma}^c \boldsymbol{n}_{it}. \tag{26}$$

where $\boldsymbol{\Gamma}^c$ is the lower triangular Cholesky decomposition of $\boldsymbol{\Sigma}^c$, i.e. $\boldsymbol{\Gamma}^c\boldsymbol{\Gamma}^{c'} = \boldsymbol{\Sigma}_c$. Appropriate re-ordering of $\boldsymbol{\epsilon}_{it}^{rc}$ yields $\boldsymbol{\epsilon}_{it}^r$.

3.4 Numerical illustration

This subsection illustrates the construction of the three simulators using a numerical example. Suppose the model is a stylized version of the canonical Keane and Wolpin (1994) model. As described above, in that model $N = 4$, $\omega = (1, 1, 0, 0)$, so that a payoff is only observed for the first two choice. For simplicity, assume that

$$\Phi_n(E_{nit}) = \begin{cases} \exp(E_{nit}) & \text{for } n = 1, 2 \\ 1 + E_{nit} & \text{for } n = 3, 4, \end{cases} \tag{27}$$

$\Psi_n = 0$ and $\Sigma = I$. Further suppose that the standard deviation of the measurement error in wages is $\sigma_\eta = 1$.

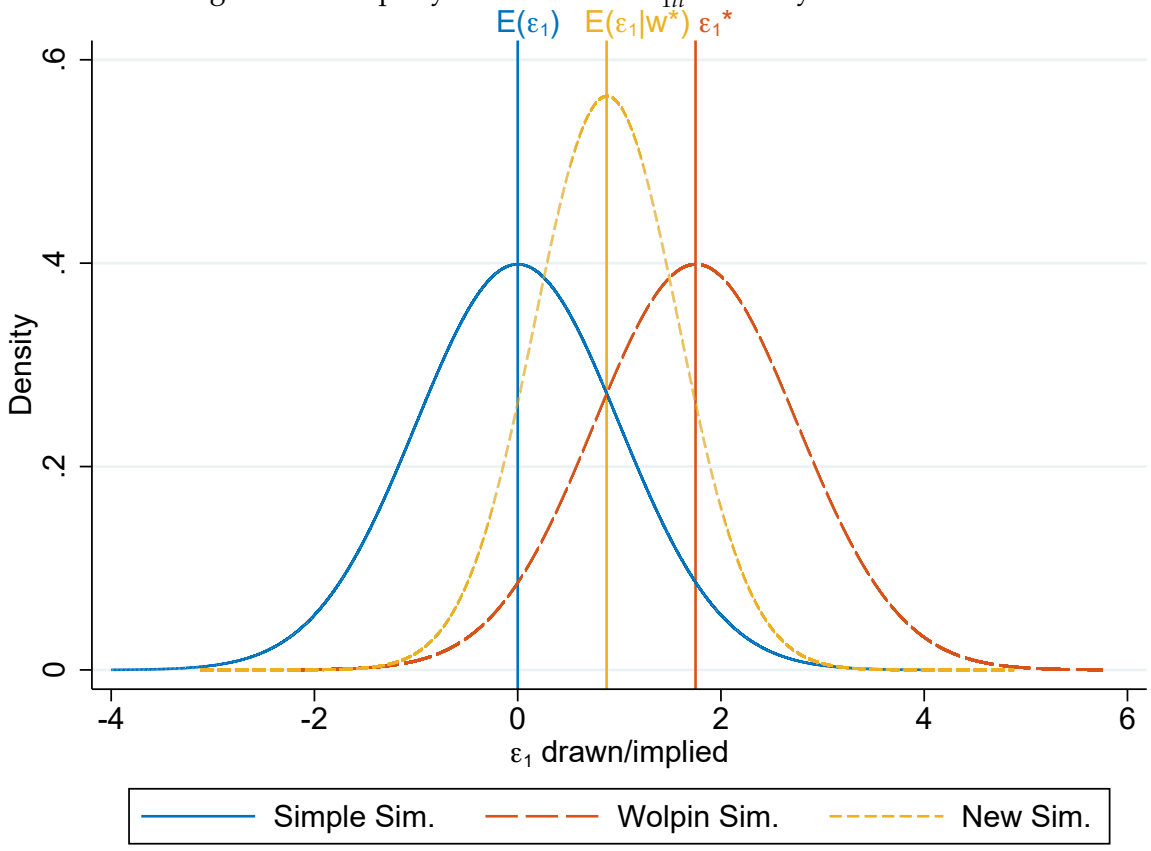
Now consider the task of simulating the joint density $f_{\boldsymbol{D},W^*}(\boldsymbol{d}_{it}, w_{it}^*)$ for a single pair of

[12]One way to obtain this expression is to use results from Bayesian statistics. ϵ_{cit} can be interpreted as the unknown mean of a normal distribution with known variance σ_η^2, and ϵ_{cit}^* as a single draw from that distribution. Hence, in Bayesian terms, $\phi_E(\epsilon_{cit})$ is the prior, $\phi_{E_c^*|E}(\epsilon_{cit}^*|\epsilon_{cit})$ is the likelihood function, $\phi_{E_c^*}(\epsilon_{cit}^*)$ is the unconditional probability of the data, and $\phi_{E|E_c^*}(\epsilon_{cit}|\epsilon_{cit}^*)$ is the posterior. It is known that when the prior is Normal with mean 0 and variance σ_c^2, and the likelihood is Normal with known variance σ_η^2, the posterior distribution of an unknown mean ϵ_{cit} given a single observation ϵ_{cit}^* is given by (24) (see e.g. Gelman et al., 2013, p. 39f).

observations, namely $\boldsymbol{d}_{it} = (1,0,0,0)$ and $w_{it}^* = 5.735$. In practice, the three simulators have much in common. All are formed as averages over R simulated densities. The simulated densities are in all cases based on drawing from four random shocks, and in each case the structural shocks ϵ_{2it}, ϵ_{3it} and ϵ_{4it} are drawn from their unconditional distribution (the standard normal distribution).

The crucial difference between the simulators concerns ϵ_{1it}. For the Simple Simulator, ϵ_{1it} is drawn from its unconditional distribution (the standard normal distribution). For the Wolpin Simulator, ϵ_{1it} is not directly drawn at all, but instead the measurement error η_{it} is drawn from its unconditional distribution (in the present case, also the standard normal distribution). Then the implied value of ϵ_{1it} can be obtained according to (13) and is distributed $N(\log(w_{it}^*), 1)$. For the new simulator, ϵ_{1it} is drawn from its distribution conditional on w_{it}^* given in (24), which in this case is $N\left(\frac{\log(w_{it}^*)}{2}, \frac{1}{2}\right)$.

Figure 1: Exemplary distribution of ϵ_{1it}^r in the stylized model



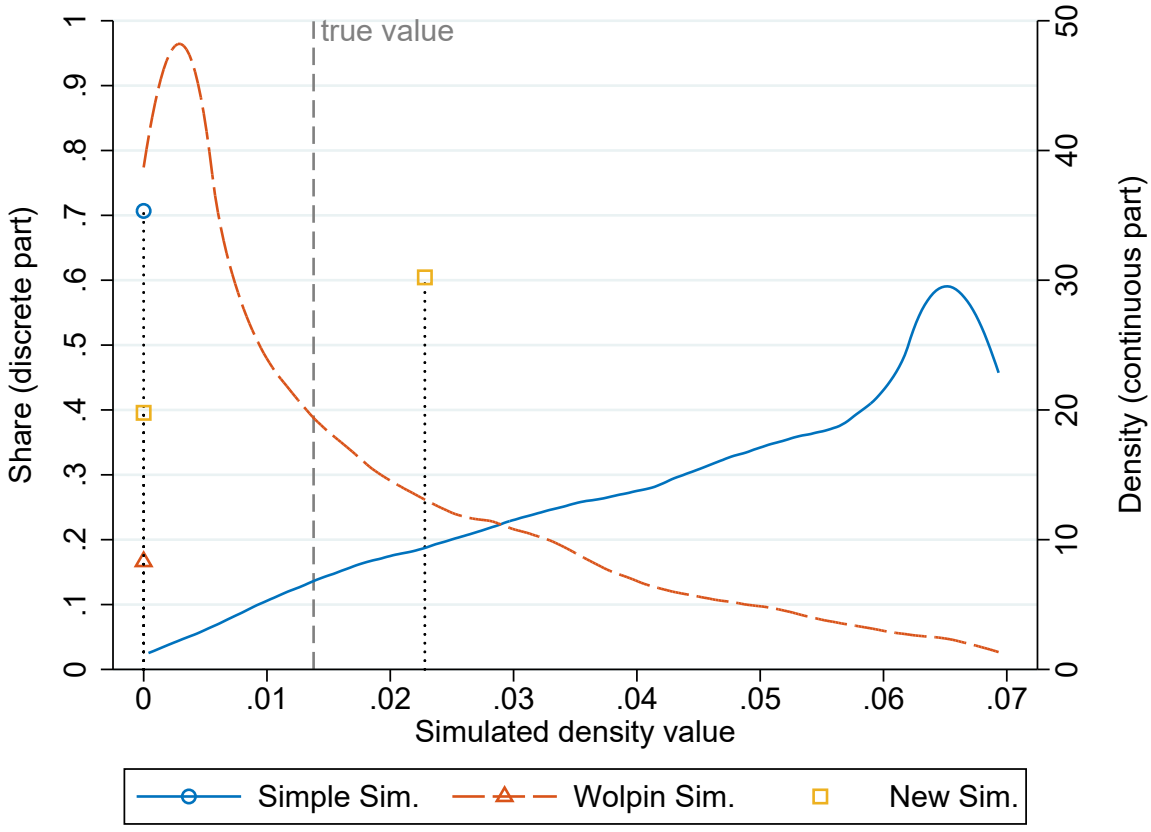
Notes. Distributions for $\boldsymbol{d}_{it} = (1,0,0,0)$ and $w_{it}^* = 5.735$.

This is illustrated in Figure 1. The Simple Simulator ignores w_{it}^* in simulation and draws

ϵ^r_{1it} from the unconditional distribution of ϵ_{1it} centred around $\mathrm{E}(\epsilon_{1it})$. The Wolpin Simulator implicitly sets ϵ^r_{1it} to values drawn symmetrically around $\epsilon^*_{1it} = \log(w^*_{1it})$, ignoring its unconditional distribution. The new simulator provides a compromise between these two approaches, with the ϵ^r_{1it} drawn from the distribution of ϵ_{1it} conditional on w^*_{it} obtained using Bayes' theorem. Bayes' theorem makes it possible to combine the unconditional ('prior') density of ϵ_{1it} with the information from observing w^*_{it} to obtain the conditional ('posterior') density.

In the stylized model as specified, the conditional mean $E(\epsilon_{1it}|w^*_{1it})$ is exactly halfway between $\mathrm{E}(\epsilon_{1it})$ and ϵ^*_{1it} because the variances of ϵ_{1it} and η_{it} are the same. If the relative standard deviation of the measurement error was large, the distribution of ϵ_{1it} conditional on w^*_{1it} would be close to the unconditional distribution, as w^*_{it} would contain little information. If the relative standard deviation of the measurement error was small, the conditional distribution would be close to the distribution of ϵ^r_{1it} implied by the Wolpin Simulator.

Figure 2: Exemplary distribution of simulated density values in the stylized model



Notes. Distributions of density values for $\boldsymbol{d}_{it} = (1, 0, 0, 0)$ and $w^*_{it} = 5.735$, calculated over $R = 10,000$ draws. The densities for the continuous part are kernel densities estimated using an Epanechnikov kernel with the bandwidth selected using Silverman's Rule.

Figure 2 shows the distribution of simulated density values obtained for each of $R = 10,000$

sets of draws when simulating $f_{\boldsymbol{D},W^*}(\boldsymbol{d}_{it}, w^*_{it})$ according to the different simulators. The distribution of the simulated density values is a mixture of discrete and continuous for the Simple and the Wolpin Simulator, as they both are products of an indicator function (which leads to a positive probability of zero) and a continuous density value that is a function of the simulation draws. The distribution of the simulated density values is a discrete distribution with two mass points, as each density value is the product of an indicator function and the positive constant $f_{V^*(\boldsymbol{d})}(w^*_{it})$, which does not depend on the simulation draws. The average simulated density value of all three simulators is very near the true density value, which is $f_{\boldsymbol{D},W^*}(\boldsymbol{d}_{it}, w^*_{it}) \approx 0.014$.

Since all three simulators are sample averages of simulated density values, the efficiency of each simulator is inversely proportional to the variance of the simulated density values. As illustrated in Figure 2, a drawback of the Simple Simulator in the given case is that in more than 70% of simulations, the simulated density value is zero, as the simulated shocks do not lead to the observed option being chosen. This is counterbalanced by density values that are on average much higher than the true value when the simulated choice does align with the observed choice. This combination leads to high variances and thus inefficiency. For the Wolpin Simulator, simulated choices line up with the observed choice in more than 80% of cases given the high observed wage, but a lot of the simulated densities are far lower than the true value, as draws often imply implausibly high values of ϵ_{1it}. Again this is counterbalanced by a few draws with much higher density values than the true value. The variance of simulated density values is lower for the new simulator, as a relatively low share of simulated densities is zero, and most are close to the true value.

3.5 Monte Carlo evidence: simulation performance

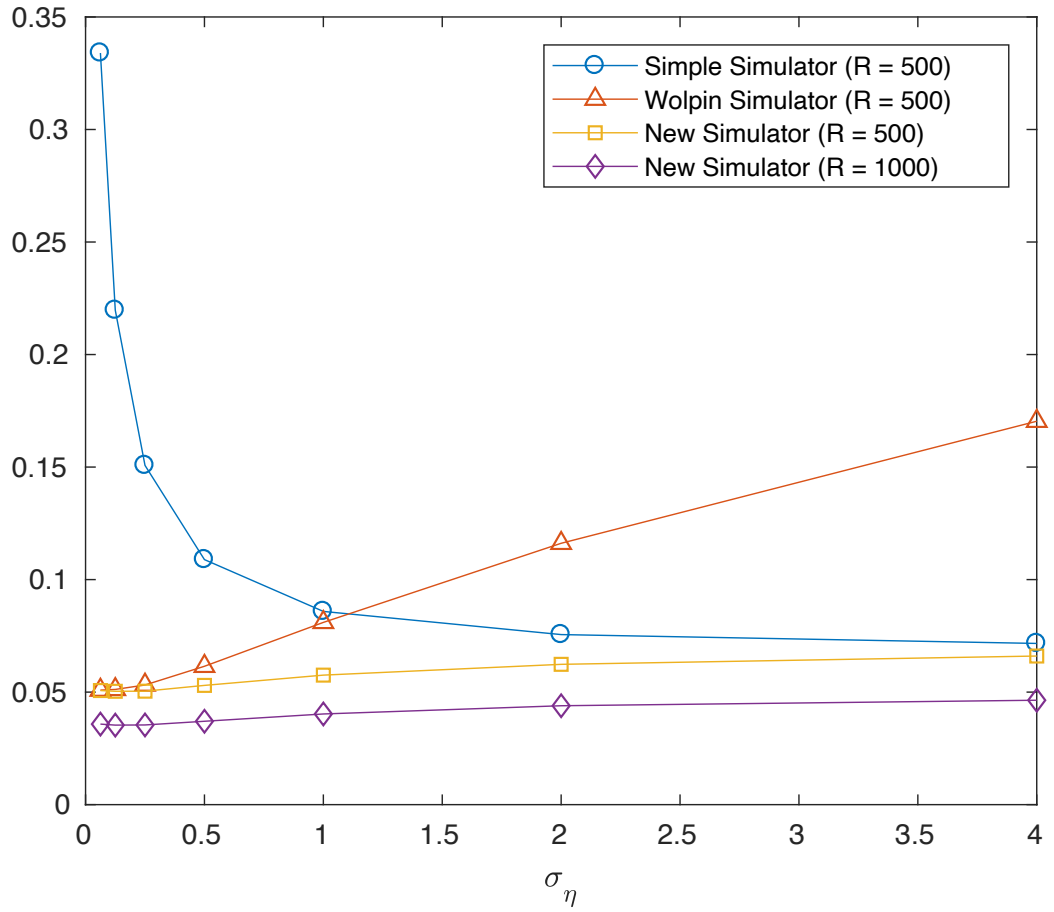
This subsection presents Monte Carlo evidence of the performance of the different simulators, using the model of 3.4 as an exemplar. Various values of the (relative) standard deviation of the measurement error in wages σ_η are considered, as this parameter crucially affects the performance of different simulators.

First, the model was used to generate simulated data. The likelihood contribution of the resulting observations was then evaluated using the three different simulators. $I = 20$ pairs of choices and observed payoffs were evaluated by simulation. For each pair, the three simulators were calculated $S = 5000$ times using different random draws. Each time, $R = 500$ draws were taken from, respectively, the unconditional distribution of $\boldsymbol{\epsilon}_{it}$, the unconditional distribution

of $\tilde{\boldsymbol{\epsilon}}_{it}$, and the distribution of $\boldsymbol{\epsilon}_{it}$ conditional on w_{it}^*. For each pair of observations and each simulator, the sample standard deviation of $\log(\hat{f})$, the logarithm of the simulated density, was calculated. Then the mean was taken across I observations. In addition, the same exercise was performed for the new simulator with $R = 1000$ draws.

The mean standard deviation of $\log(\hat{f})$ was chosen as a measure because it has two intuitive interpretations. First, it is an unbiased estimator of the root mean squared error of the log-likelihood contribution for a given observation. Second, for small values, it approximates the standard percentage error in an observation's simulated likelihood contribution.

Figure 3: Mean standard deviation of $\log(\hat{f})$, baseline case



Notes. All results shown are means for $I = 20$ pairs of $\boldsymbol{d}_{it}$ and w_{it}^*. In each case, the sample standard deviation is taken over $S = 5000$ evaluations.

Figure 3 shows the results. The new simulator performs as well as either simulator from the literature for all values of the standard deviation of the measurement error σ_η. The new simulator performs as well as the Wolpin Simulator for small values of σ_η and roughly as well

as the Simple Simulator for large σ_η. For intermediate values of σ_η, the new simulator clearly outperforms both the Wolpin Simulator and the Simple Simulator.

These results are reasonable given the way the three simulators are constructed. The reason that the new simulator outperforms the Simple Simulator is that w_{it}^* contains information about $\boldsymbol{\epsilon}_{it}$, which the new simulator does and the Simple Simulator does not take into account in sampling. However, as $\sigma_\eta \to \infty$, the observed payoff becomes uninformative about the true payoff, so the performance of the two simulators converges.

The Wolpin Simulator, on the other hand, does take into account the information contained in the observed payoff, but uses it in a suboptimal way. Sampling from the unconditional distribution of the measurement error does not take into account that some values for the true payoff are more likely than others *ex ante*. However, if σ_η is very small, this issue is not relevant, as then the randomness introduced by sampling from the distribution of the measurement error is negligible. Hence the performance of the Wolpin Simulator converges to that of the new simulator as $\sigma_\eta \to 0$.[13]

Table 1: Baseline: average computation time for $\hat{f}_{\boldsymbol{D},W^*}(\boldsymbol{d}_{it}, w_{it}^*)$

	Simple Sim. $R = 500$	Wolpin Sim. $R = 500$	New Sim. $R = 500$	New Sim. $R = 1000$
Time (in ms)	0.238	0.279	0.199	0.256

As Table 1 shows, the new simulator is also substantially faster to compute.[14] The reason is that the density of the observed payoff only has to be evaluated once as opposed to R times for the other two simulators. If the sample size is increased to bring computation time in line with the other two Simulators, the new simulator is more accurate for all values of σ_η. Notably, if $\sigma_\eta = 1$, the mean standard deviation of $\log(\hat{f})$ is less than half as large as for either simulator from the literature.

[13] A further disadvantage of the Wolpin Simulator is that it does not take into account information contained in the observed payoff about the structural shocks corresponding to *non-chosen* options. This is irrelevant in the present case only because the covariance matrix of the structural shocks Σ is diagonal, so the observed payoff contains no information about other structural shocks. Results for an alternative version of the stylized model with different payoffs and a different correlation structure are presented in Appendix D.

[14] Computation time was calculated excluding random draws, but including all other calculations. The rationale for excluding random draws is that they only need to be generated once in actual estimation. All computation was performed on a 1.6GHz Intel Pentium Quadcore CPU using Matlab 2017a for Linux.

4 Unobserved state variables

Allowing for unobserved state variables that vary over time poses a challenge for the Maximum Likelihood estimation of dynamic discrete choice models, as it makes it necessary to evaluate all periods of an individual's likelihood contribution jointly. This usually requires the simulation of an integral of order NT. For example, suppose that instead of serial independence, $\boldsymbol{E}_{i1} = \boldsymbol{U}_{i1}$ and $\boldsymbol{E}_{it} = \rho \boldsymbol{E}_{i,t-1} + \boldsymbol{U}_{it}$, where $\boldsymbol{U}_{it}$ is distributed $N(0, \boldsymbol{\Sigma})$ and $\rho \neq 0$.[15]

In this section, I present a generalization of the new simulator that can accommodate this case. Both the Simple Simulator and the Wolpin Simulator can be generalized analogously. In order to make this generalization computationally feasible for panels of more than a few periods, I introduce a resampling step as in the 'bootstrap filter' of Gordon, Salmond and Smith (1993). This Sequential Monte Carlo approach has two main advantages. First, unlike the commonly used method of Keane and Sauer (2010), it does not require a change in the structural model, so the resulting bias is avoided. Second, it provides acceptable accuracy even when the number of periods T is large.[16]

4.1 Generalizing the simulators

Define $\boldsymbol{U}_i = (\boldsymbol{U}_{i1}, \ldots, \boldsymbol{U}_{iT})$ to be the NxT matrix of innovations to the structural disturbances and $\boldsymbol{u}_i$ to be a particular realization of that matrix. As a first step, I write the likelihood contribution of each individual as

$$f_{\{\boldsymbol{D}_t, W_t^*\}_{t=1}^T}(\{\boldsymbol{d}_{it}, w_{it}^*\}_{t=1}^T) = \int f_{\boldsymbol{D}_1, V_1^*(\boldsymbol{d}_1), \boldsymbol{U}_1}(\boldsymbol{d}_{i1}, w_{i1}^*, \boldsymbol{u}_{i1}) \prod_{t=2}^{T} f_{\boldsymbol{D}_t, V_t^*(\boldsymbol{d}_t), \boldsymbol{U}_t | \{\boldsymbol{D}_s, V_s^*(\boldsymbol{d}_s), \boldsymbol{U}_s\}_{s=1}^{t-1}}(\boldsymbol{d}_t, w_{it}^*, \boldsymbol{u}_{it} | \{\boldsymbol{d}_{is}, w_{is}^*, \boldsymbol{u}_{is}\}_{s=1}^{t-1}) d\boldsymbol{u}_i. \quad (28)$$

Further conditioning and simplifying yields

$$f_{\{\boldsymbol{D}_t, W_t^*\}_{t=1}^T}(\{\boldsymbol{d}_{it}, w_{it}^*\}_{t=1}^T) = \int f_{V_1^*(\boldsymbol{d}_1)}(w_{i1}^*) P(\boldsymbol{D}_{i1} = \boldsymbol{d}_{i1} | \boldsymbol{U}_{i1} = \boldsymbol{u}_{i1}) \phi_{\boldsymbol{U}_1 | V_1^*(\boldsymbol{d}_1)}(\boldsymbol{u}_{i1} | w_{i1}^*) \prod_{t=2}^{T} f_{V_t^*(\boldsymbol{d}_t) | \{\boldsymbol{U}_s\}_{s=1}^{t-1}}(w_{it}^* | \{\boldsymbol{u}_{is}\}_{s=1}^{t-1}) P(\boldsymbol{D}_{it} = \boldsymbol{d}_{it} | \{\boldsymbol{U}_{is} = \boldsymbol{u}_{is}\}_{s=1}^{t}) \phi_{\boldsymbol{U}_t | V_t^*(\boldsymbol{d}_t)}(\boldsymbol{u}_{it} | w_{it}^*) d\boldsymbol{u}_i \quad (29)$$

[15] For instance, in the context of the Keane and Wolpin (1994) model, this would allow for persistence in shocks to wages and preferences.

[16] See Creal (2012) for other uses of Sequential Monte Carlo methods in economics, and Blevins (2016) for a discussion of the application of these methods to the estimation of dynamic discrete choice models.

Hence a consistent simulator is

$$\hat{f}^N(\{\boldsymbol{d}_{it}\}_{t=1}^T, \{w_{it}^*\}_{t=1}^T) =$$
$$\frac{f_{V_1^*(\boldsymbol{d}_1)}(w_{i1}^*)}{R} \sum_{r=1}^{R} P(\boldsymbol{D}_{i1} = \boldsymbol{d}_{i1} | \boldsymbol{U}_{i1} = \boldsymbol{u}_{i1}^r) \prod_{t=2}^{T} f_{V_t^*(\boldsymbol{d}_t)|\{\boldsymbol{U}_s\}_{s=1}^{t-1}}(w_{it}^* | \{\boldsymbol{u}_{is}^r\}_{s=1}^{t-1}) P(\boldsymbol{D}_{it} = \boldsymbol{d}_{it} | \{\boldsymbol{U}_{is} = \boldsymbol{u}_{is}^r\}_{s=1}^{t}) \tag{30}$$

where each element of the sequence $\{\boldsymbol{u}_{it}^r\}_{t=1}^T$ is drawn conditional on the observed wage in the same period using the approach developed in section 3.3. The expressions for the choice probability and the observed wage are evaluated as in that section.

The analogous expression for the Simple Simulator is

$$\hat{f}^S_{\{\boldsymbol{D}_t, W_t^*\}_{t=1}^T}(\{\boldsymbol{d}_{it}, w_{it}^*\}_{t=1}^T) = \frac{1}{R} \sum_{r=1}^{R} \prod_{t=1}^{T} f_{W^*|\boldsymbol{U}}(w_{it}^* | \boldsymbol{u}_i^r) P(\boldsymbol{D}_{it} = \boldsymbol{d}_{it} | \boldsymbol{U}_i = \boldsymbol{u}_i^r) \tag{31}$$

where the innovations $\boldsymbol{u}_i^r$ are drawn from their unconditional distribution. The analogous expression for the Wolpin simulator is

$$\hat{f}^W_{\{\boldsymbol{D}_t, W_t^*\}_{t=1}^T}(\{\boldsymbol{d}_{it}, w_{it}^*\}_{t=1}^T) =$$
$$\frac{1}{R} \sum_{r=1}^{R} f_{W_1^*|\tilde{\boldsymbol{U}}_1}(w_{i1}^* | \tilde{\boldsymbol{u}}_{i1}^r) P(\boldsymbol{D}_{i1} = \boldsymbol{d}_{i1} | \tilde{\boldsymbol{U}}_{i1} = \tilde{\boldsymbol{u}}_{i1}^r, W_{i1}^* = w_{i1}^*)$$
$$\prod_{t=2}^{T} f_{W^*|\tilde{\boldsymbol{U}}}(w_{it}^* | \{\tilde{\boldsymbol{u}}_{is}^r\}_{s=1}^{t}, \{w_{is}^*\}_{s=1}^{t-1}) P(\boldsymbol{D}_{it} = \boldsymbol{d}_{it} | \{\tilde{\boldsymbol{U}}_{is} = \tilde{\boldsymbol{u}}_{is}^r\}_{s=1}^{t}, \{W_{is}^* = w_{is}^*\}_{s=1}^{t}) \tag{32}$$

where the innovations $\tilde{\boldsymbol{u}}_i^r$ are drawn from their unconditional distribution.

Unfortunately, all of these simulators perform poorly if the number of periods T is large. One reason is that for any feasible R, some choice histories in the data may not match any of R simulated choice histories, leading to a simulated likelihood of zero. Resampling helps to avoid this problem.

4.2 Resampling

Observing likelihood values of zero is only one symptom of a broader problem. As the number of periods T increases, the sums in (30), (31) and (32) will come to be dominated by a single summand indexed by r.[17] As shown in Gordon, Salmond and Smith (1993), resampling

[17] A closely related problem is known in the Sequential Monte Carlo literature as *weight degeneracy* (see e.g. Creal, 2012).

provides a solution to this problem. In contrast to standard methods from the discrete choice literature, the *Sequential Monte Carlo* methods based on this idea do not introduce any bias in the simulated likelihood.[18]

The starting point for these methods is the *prediction decomposition*:

$$f_{\{\boldsymbol{D}_t, W_t^*\}_{t=1}^T}(\{\boldsymbol{d}_{it}, w_{it}^*\}_{t=1}^T) = \prod_{t=1}^T f_{\boldsymbol{D}_t, V_t^*(\boldsymbol{d}_t)|\{\boldsymbol{D}_s, V_s^*(\boldsymbol{d}_s)\}_{s=1}^{t-1}}(\boldsymbol{d}_t, w_{it}^*|\{\boldsymbol{d}_{is}, w_{is}^*\}_{s=1}^{t-1}). \tag{33}$$

Then each factor of this product can be written as

$$f_{\boldsymbol{D}_t, V_t^*(\boldsymbol{d}_t)|\{\boldsymbol{D}_s, V_s^*(\boldsymbol{d}_s)\}_{s=1}^{t-1}}(\boldsymbol{d}_t, w_{it}^*|\{\boldsymbol{d}_{is}, w_{is}^*\}_{s=1}^{t-1}) = \int f_{\boldsymbol{D}_t, V_t^*(\boldsymbol{d}_t)|\boldsymbol{E}_{it}}(\boldsymbol{d}_t, w_{it}^*|\boldsymbol{\epsilon}_{it}) f_{\boldsymbol{E}_{it}|\boldsymbol{D}_t, V_t^*(\boldsymbol{d}_t)}(\boldsymbol{\epsilon}_{it}|\{\boldsymbol{d}_{is}, w_{is}^*\}_{s=1}^{t-1}) d\boldsymbol{\epsilon}_{it} \tag{34}$$

or alternatively — to align with the new simulator — as

$$\begin{aligned} f_{\boldsymbol{D}_t, V_t^*(\boldsymbol{d}_t)|\{\boldsymbol{D}_s, V_s^*(\boldsymbol{d}_s)\}_{s=1}^{t-1}}(\boldsymbol{d}_t, w_{it}^*|\{\boldsymbol{d}_{is}, w_{is}^*\}_{s=1}^{t-1}) = \\ \iint P(\boldsymbol{d}_t|\boldsymbol{\epsilon}_{it}) f_{V_t^*(\boldsymbol{d}_t)|\boldsymbol{E}_{i,t-1}}(w_{it}^*|\boldsymbol{\epsilon}_{i,t-1}) f_{\boldsymbol{E}_{it}|\boldsymbol{E}_{i,t-1}, V_t^*(\boldsymbol{d}_t)}(\boldsymbol{\epsilon}_{it}|\boldsymbol{\epsilon}_{i,t-1}, w_{it}^*) \\ f_{\boldsymbol{E}_{i,t-1}|\{\boldsymbol{D}_s, V_s^*(\boldsymbol{d}_s)\}_{s=1}^{t-1}}(\boldsymbol{\epsilon}_{i,t-1}|\{\boldsymbol{d}_{is}, w_{is}^*\}_{s=1}^{t-1}) d\boldsymbol{\epsilon}_{it} d\boldsymbol{\epsilon}_{i,t-1}. \end{aligned} \tag{35}$$

Hence a consistent and unbiased simulator of $f_{\boldsymbol{D}_t, V_t^*(\boldsymbol{d}_t)|\{\boldsymbol{D}_s, V_s^*(\boldsymbol{d}_s)\}_{s=1}^{t-1}}(\boldsymbol{d}_t, w_{it}^*|\{\boldsymbol{d}_{is}, w_{is}^*\}_{s=1}^{t-1})$ is

$$\hat{f}^N_{\boldsymbol{D}_t, V_t^*(\boldsymbol{d}_t)|\{\boldsymbol{D}_s, V_s^*(\boldsymbol{d}_s)\}_{s=1}^{t-1}}(\boldsymbol{d}_t, w_{it}^*|\{\boldsymbol{d}_{is}, w_{is}^*\}_{s=1}^{t-1}) = \frac{1}{R}\sum_{r=1}^R P(\boldsymbol{d}_t|\boldsymbol{\epsilon}_{it}^r) f_{V_t^*(\boldsymbol{d}_t)|\boldsymbol{E}_{i,t-1}}(w_{it}^*|\boldsymbol{\epsilon}_{i,t-1}^r) \tag{36}$$

where $\boldsymbol{\epsilon}_{i,t-1}^r$ is drawn from the distribution of $\boldsymbol{E}_{i,t-1}$ conditional on the whole history of choices and observed wages $\{\boldsymbol{d}_{is}, w_{is}^*\}_{s=1}^{t-1}$, and $\boldsymbol{\epsilon}_{it}^r$ is drawn from the distribution of $\boldsymbol{E}_{it}$ conditional on $\boldsymbol{\epsilon}_{i,t-1}^r$ and w_{it}^*.

$P(\boldsymbol{d}_t|\boldsymbol{\epsilon}_{it})$ and $f_{V_t^*(\boldsymbol{d}_t)|\boldsymbol{E}_{i,t-1}}(w_{it}^*|\boldsymbol{\epsilon}_{i,t-1})$ are straightforward to evaluate, and drawing from the distribution of $\boldsymbol{E}_{it}$ conditional on $\boldsymbol{\epsilon}_{i,t-1}^r$ and w_{it}^* poses no additional difficulties compared to section 3.3. The challenge is to draw from $\boldsymbol{\epsilon}_{i,t-1}^r$ from the distribution of $\boldsymbol{E}_{i,t-1}$ conditional on the whole history of choices and observed wages $\{\boldsymbol{d}_{is}, w_{is}^*\}_{s=1}^{t-1}$. This can be accomplished using a *particle filter*.

[18] A downside of this family of methods is that the resulting simulated likelihood is not a smooth function of the parameters even if the expressions for the choice probabilities are smooth. Malik and Pitt (2011) offer a technique for smoothing the likelihood function in this case, which is especially useful when the unobserved state is univariate.

The crucial step is to get to a sample from $f_{\boldsymbol{E}_{it}|\{\boldsymbol{D}_s,V_s^*(\boldsymbol{d}_s)\}_{s=1}^t}(\boldsymbol{\epsilon}_{it}|\{\boldsymbol{d}_{is},w_{is}^*\}_{s=1}^t)$ from a sample from $f_{\boldsymbol{E}_{it}|\{\boldsymbol{D}_s,V_s^*(\boldsymbol{d}_s)\}_{s=1}^{t-1}}(\boldsymbol{\epsilon}_{it}|\{\boldsymbol{d}_{is},w_{is}^*\}_{s=1}^{t-1})$. Getting a sample from $f_{\boldsymbol{E}_{i,t}|\{\boldsymbol{D}_s,V_s^*(\boldsymbol{d}_s)\}_{s=1}^{t-1}}(\boldsymbol{\epsilon}_{it}|\{\boldsymbol{d}_{is},w_{is}^*\}_{s=1}^{t-1})$ from a sample from $f_{\boldsymbol{E}_{i,t-1}|\{\boldsymbol{D}_s,V_s^*(\boldsymbol{d}_s)\}_{s=1}^{t-1}}(\boldsymbol{\epsilon}_{i,t-1}|\{\boldsymbol{d}_{is},w_{is}^*\}_{s=1}^{t-1})$ is straightforward, because $\{\boldsymbol{d}_{is},w_{is}^*\}_{s=1}^{t-1}$ only affects $\boldsymbol{\epsilon}_{it}$ through $\boldsymbol{\epsilon}_{i,t-1}$. Hence given a draw $\boldsymbol{\epsilon}_{i,t-1}^r$ from $f_{\boldsymbol{E}_{i,t-1}|\{\boldsymbol{D}_s,V_s^*(\boldsymbol{d}_s)\}_{s=1}^{t-1}}(\boldsymbol{\epsilon}_{i,t-1}|\{\boldsymbol{d}_{is},w_{is}^*\}_{s=1}^{t-1})$, one only needs to sample from $f_{\boldsymbol{E}_{i,t}|\boldsymbol{E}_{i,t-1}}(\boldsymbol{\epsilon}_{it}|\boldsymbol{\epsilon}_{i,t-1}^r)$ to obtain a sample from $f_{\boldsymbol{E}_{i,t}|\{\boldsymbol{D}_s,V_s^*(\boldsymbol{d}_s)\}_{s=1}^{t-1}}(\boldsymbol{\epsilon}_{it}|\{\boldsymbol{d}_{is},w_{is}^*\}_{s=1}^{t-1})$.

A sample from $f_{\boldsymbol{E}_{it}|\{\boldsymbol{D}_s,V_s^*(\boldsymbol{d}_s)\}_{s=1}^{t}}(\boldsymbol{\epsilon}_{it}|\{\boldsymbol{d}_{is},w_{is}^*\}_{s=1}^{t})$ can be obtained though resampling. By Bayes' theorem,

$$f_{\boldsymbol{E}_{it}|\{\boldsymbol{D}_s,V_s^*(\boldsymbol{d}_s)\}_{s=1}^{t}}(\boldsymbol{\epsilon}_{it}|\{\boldsymbol{d}_{is},w_{is}^*\}_{s=1}^{t}) \propto$$
$$f_{\boldsymbol{D}_t,V_t^*(\boldsymbol{d}_t)|\boldsymbol{E}_{it},\{\boldsymbol{D}_s,V_s^*(\boldsymbol{d}_s)\}_{s=1}^{t-1}}(\boldsymbol{d}_{it},w_{it}^*|\boldsymbol{\epsilon}_{it},\{\boldsymbol{d}_{is},w_{is}^*\}_{s=1}^{t-1}) f_{\boldsymbol{E}_{it}|\{\boldsymbol{D}_s,V_s^*(\boldsymbol{d}_s)\}_{s=1}^{t-1}}(\boldsymbol{\epsilon}_{it}|\{\boldsymbol{d}_{is},w_{is}^*\}_{s=1}^{t-1}). \quad (37)$$

As, in addition, $f_{\boldsymbol{E}_{it}|\{\boldsymbol{D}_s,V_s^*(\boldsymbol{d}_s)\}_{s=1}^{t}}(\boldsymbol{\epsilon}_{it}|\{\boldsymbol{d}_{is},w_{is}^*\}_{s=1}^{t})$ is absolutely continuous with respect to $f_{\boldsymbol{E}_{it}|\{\boldsymbol{D}_s,V_s^*(\boldsymbol{d}_s)\}_{s=1}^{t-1}}(\boldsymbol{\epsilon}_{it}|\{\boldsymbol{d}_{is},w_{is}^*\}_{s=1}^{t-1})$, a sample from the former can be obtained given a sample from the latter using a *weighted bootstrap* . For each draw from $f_{\boldsymbol{E}_{it}|\{\boldsymbol{D}_s,V_s^*(\boldsymbol{d}_s)\}_{s=1}^{t-1}}(\boldsymbol{\epsilon}_{it}|\{\boldsymbol{d}_{is},w_{is}^*\}_{s=1}^{t-1})$, one can calculate

$$\pi_{it}^r = \frac{\omega_{it}^r}{\sum_{r=1}^R \omega_{it}^r}$$

where

$$\omega_{it}^r = f_{\boldsymbol{D}_t,V_t^*(\boldsymbol{d}_t)|\boldsymbol{E}_t,\{\boldsymbol{D}_s,V_s^*(\boldsymbol{d}_s)\}_{s=1}^{t-1}}(\boldsymbol{d}_{it},w_{it}^*|\boldsymbol{\epsilon}_{it}^r,\{\boldsymbol{d}_{is},w_{is}^*\}_{s=1}^{t-1}) = f_{\boldsymbol{D}_t,V_t^*(\boldsymbol{d}_t)|\boldsymbol{E}_t}(\boldsymbol{d}_{it},w_{it}^*|\boldsymbol{\epsilon}_{it}^r).$$

Then a sample from the discrete distribution over the set of draws $\{\boldsymbol{\epsilon}_{it}^r\}_{r=1}^R$ from $f_{\boldsymbol{E}_{it}|\{\boldsymbol{D}_s,V_s^*(\boldsymbol{d}_s)\}_{s=1}^{t-1}}(\boldsymbol{\epsilon}_{it}|\{\boldsymbol{d}_{is},w_{is}^*\}_{s=1}^{t-1})$, placing probability mass π_{it}^r on $\boldsymbol{\epsilon}_{it}^r$, will tend in distribution to $f_{\boldsymbol{E}_{it}|\{\boldsymbol{D}_s,V_s^*(\boldsymbol{d}_s)\}_{s=1}^{t}}(\boldsymbol{\epsilon}_{it}|\{\boldsymbol{d}_{is},w_{is}^*\}_{s=1}^{t})$ as $R \to \infty$ (for a proof, see Smith and Gelfand, 1992).

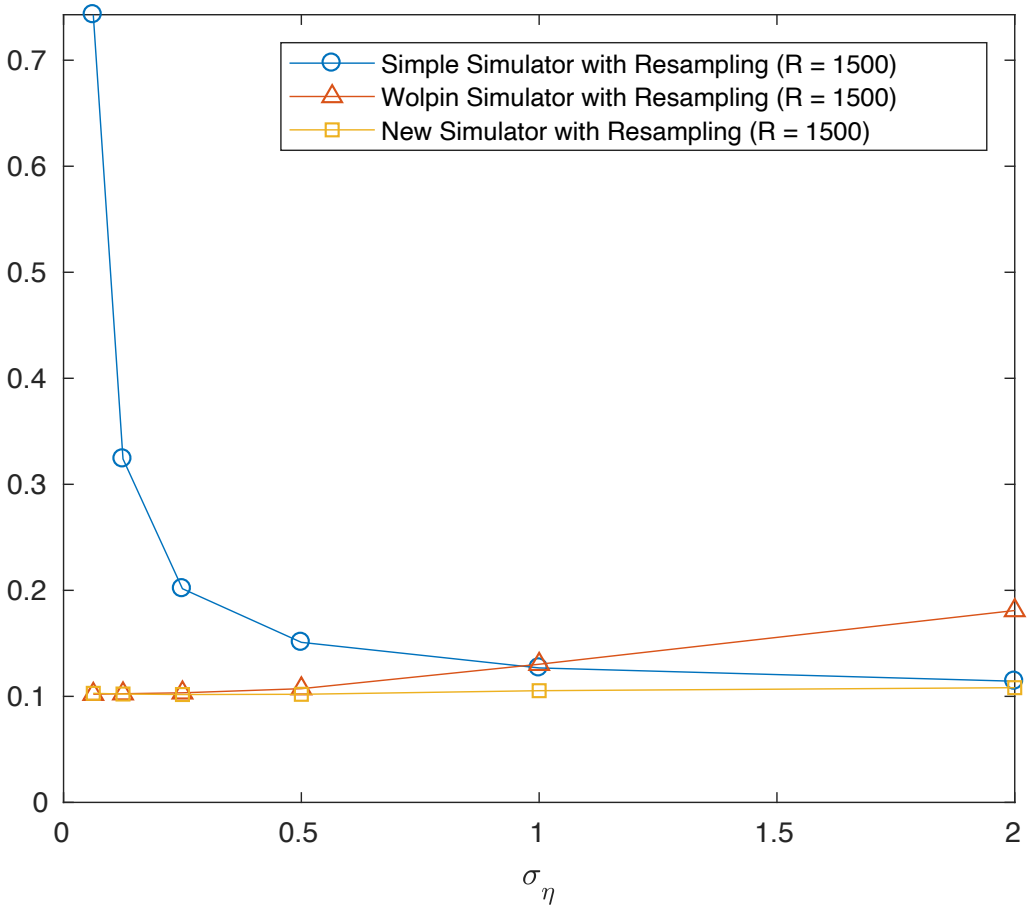
This method of drawing from $f_{\boldsymbol{E}_{it}|\{\boldsymbol{D}_s,V_s^*(\boldsymbol{d}_s)\}_{s=1}^{t}}(\boldsymbol{\epsilon}_{it}|\{\boldsymbol{d}_{is},w_{is}^*\}_{s=1}^{t})$, is the *bootstrap filter* originally proposed by Gordon, Salmond and Smith (1993). Using this filter, it is possible to evaluate the conditional likelihood expressions above: (33) for the Simple Simulator and (34) for the new simulator. However, for the new simulator, it is advisable to go one step further and use the same importance sampling technique in solving the filtering problem.

In particular, instead of drawing from $f_{\boldsymbol{E}_{it}|\{\boldsymbol{D}_s,V_s^*(\boldsymbol{d}_s)\}_{s=1}^{t-1}}(\boldsymbol{\epsilon}_{it}|\{\boldsymbol{d}_{is},w_{is}^*\}_{s=1}^{t-1})$ in the weighted bootstrap, one can draw from $f_{\boldsymbol{E}_{it}|\{\boldsymbol{D}_s,V_s^*(\boldsymbol{d}_s)\}_{s=1}^{t-1},V_t^*(\boldsymbol{d}_t)}(\boldsymbol{\epsilon}_{it}|\{\boldsymbol{d}_{is},w_{is}^*\}_{s=1}^{t-1},w_{it}^*)$, which can be accomplished in practice by drawing from $f_{\boldsymbol{E}_{i,t}|\boldsymbol{E}_{i,t-1},V_t^*(\boldsymbol{d}_t)}(\boldsymbol{\epsilon}_{it}|\boldsymbol{\epsilon}_{i,t-1}^r,w_{it}^*)$. As $f_{\boldsymbol{E}_{it}|\{\boldsymbol{D}_s,V_s^*(\boldsymbol{d}_s)\}_{s=1}^{t}}(\boldsymbol{\epsilon}_{it}|\{\boldsymbol{d}_{is},w_{is}^*\}_{s=1}^{t})$

is also absolutely continuous with respect to $f_{\boldsymbol{E}_{it}|\{\boldsymbol{D}_s,V_s^*(\boldsymbol{d}_s)\}_{s=1}^{t-1},V_t^*(\boldsymbol{d}_t)}(\boldsymbol{\epsilon}_{it}|\{\boldsymbol{d}_{is},w_{is}^*\}_{s=1}^{t-1},w_{it}^*)$, all that changes compared to the *bootstrap filter* is the weight ω_{it}^r, which becomes

$$\begin{aligned}\tilde{\omega}_{it}^r &= \frac{f_{\boldsymbol{E}_{it}|\{\boldsymbol{D}_s,V_s^*(\boldsymbol{d}_s)\}_{s=1}^{t-1}}(\boldsymbol{\epsilon}_{it}|\{\boldsymbol{d}_{is},w_{is}^*\}_{s=1}^{t-1})}{f_{\boldsymbol{E}_{it}|\{\boldsymbol{D}_s,V_s^*(\boldsymbol{d}_s)\}_{s=1}^{t-1},V_t^*(\boldsymbol{d}_t)}(\boldsymbol{\epsilon}_{it}|\{\boldsymbol{d}_{is},w_{is}^*\}_{s=1}^{t-1},w_{it}^*)}\omega_{it}^r \\ &= P(\boldsymbol{d}_t|\boldsymbol{\epsilon}_{it}^r)f_{V_t^*(\boldsymbol{d}_t)|\boldsymbol{E}_{i,t-1}}(w_{it}^*|\boldsymbol{\epsilon}_{i,t-1}^r).\end{aligned}$$

Figure 4: Mean standard deviation of $\log(\hat{f})$, Unobserved State Variables



Notes. All results shown are means for $I = 20$ sequences of $\{\boldsymbol{d}_{it}\}_{t=1}^T$ and $\{w_{it}^*\}_{t=1}^T$, where $T = 5$. In each case, the sample standard deviation is taken over $S = 1000$ evaluations.

This improves the efficiency of the procedure for the same reason that the new simulator outperforms the Simple Simulator: The observed wage w_{it}^* is taken into account at the sampling stage in each period. Conveniently, using this alternative particle filter, an estimate for

the conditional likelihood in (34) very nearly emerges as a by-product, as

$$\hat{f}^{N}_{\boldsymbol{D}_t, V_t^*(\boldsymbol{d}_t)|\{\boldsymbol{D}_s, V_s^*(\boldsymbol{d}_s)\}_{s=1}^{t-1}}(\boldsymbol{d}_t, w_{it}^*|\{\boldsymbol{d}_{is}, w_{is}^*\}_{s=1}^{t-1}) = \frac{1}{R}\sum_{r=1}^{R}\tilde{\omega}_{it}^r. \tag{38}$$

It is worth noting that resampling need not be performed every period; different schemes are possible corresponding to different ways of decomposing the joint likelihood. A general algorithm for computing the new simulator using the method outlined above, which can accommodate arbitrary patterns of resampling, is outlined in Appendix B. Analogous procedures can be followed for the Simple Simulator and the Wolpin Simulator.

4.3 Monte Carlo evidence: simulation performance

As shown in Figure 4, similar results to the baseline case obtain if unobserved state variables are introduced by way of serial correlation in the structural errors and a resampling step is introduced to simulate the likelihood. The new simulator performs at least as well as the alternative simulators, holding sample size constant. Differences in accuracy between different simulators appear somewhat smaller, because choice histories now include periods in which no payoff is observed, so that no difference between the simulators arises.

Table 2 shows that computation times for the different simulators are comparable. The resampling step is computationally expensive: Acceptable accuracy in likelihood simulation is only achieved at the cost of much higher computation time than in the baseline case.

Table 2: Unobserved state variables: average computation time for $\hat{f}(\{\boldsymbol{d}_{it}^*\}_{t=1}^{T}, \{w_{it}^*\}_{t=1}^{T})$

	Simple Sim. Res $R = 1500$	Wolpin Sim. Res. $R = 1500$	New Sim. Res. $R = 1500$
Time (in ms)	29.583	30.583	30.161

5 Classification error

Classification error is a concern in many commonly used data sets. Ignoring it in estimation generally leads to inconsistent coefficient estimates (Hausman, Abrevaya and Scott-Morton, 1998). All of the likelihood simulators discussed above can accommodate classification error, but it is much simpler to take it into account in the Simple Simulator than in the Wolpin Simulator and the new simulator.

For example, suppose there is classification error so that within the subsets of choices D_o and D_u, for which a payoff is and is not observed, respectively, there is some probability that the wrong choice is recorded. Following Keane and Sauer (2010), one might adopt the following functional form for the classification error. Let

$$\boldsymbol{l}_{it} = \gamma_1 \boldsymbol{d}'_{it} + \gamma_2 \boldsymbol{d}^{*'}_{i,t-1} + \boldsymbol{\xi}_{it}. \tag{39}$$

$\boldsymbol{\xi}_{it}$ is a vector of independent random draws; γ_1 and γ_2 govern the prevalence and persistence of recording error; $\boldsymbol{d}^*_{it}$ is the observed choice. Then the maximum of $\boldsymbol{l}_{it}$ taken over the elements corresponding to the relevant subset of choices D_o or D_u determines which choice is recorded. In what follows, I assume for simplicity that $\boldsymbol{d}^*_{it} \in D_o$; all results carry over to the more general case.

If $\boldsymbol{\xi}_{it}$ is standard Gumbel distributed, this yields a Multinomial Logit model with the conditional classification probability

$$P(D^*_{nit} = 1 | \boldsymbol{D}^*_{i,t-1} = \boldsymbol{d}^*_{i,t-1}, \boldsymbol{D}_{it} = \boldsymbol{d}_{it}) = \frac{\exp(\gamma_1 d_{nit} + \gamma_2 d^*_{ni,t-1})}{\sum_{k=1}^{N_o} \exp(\gamma_1 d_{kit} + \gamma_2 d^*_{ki,t-1})} \tag{40}$$

where N_o is the number of options for which a payoff is observed.

Classification error of this kind is straightforward to accommodate using the Simple Simulator. The conditional probability of the observed choice will take the form

$$\begin{aligned} &P(\boldsymbol{D}^*_{it} = \boldsymbol{d}^*_{it} | \boldsymbol{D}^*_{i,t-1} = \boldsymbol{d}^*_{i,t-1}, \boldsymbol{E}_{it} = \boldsymbol{\epsilon}^r_{it}) \\ &\qquad = \sum_{\boldsymbol{d}_{it} \in D_o} P(D^*_{cit} = 1 | \boldsymbol{D}^*_{i,t-1} = \boldsymbol{d}^*_{i,t-1}, \boldsymbol{D}_{it} = \boldsymbol{d}_{it}) P(\boldsymbol{D}_{it} = \boldsymbol{d}_{it} | \boldsymbol{E}_{it} = \boldsymbol{\epsilon}^r_{it}) \end{aligned} \tag{41}$$

where $P(\boldsymbol{D}_{it} = \boldsymbol{d}_{it} | \boldsymbol{E}_{it} = \boldsymbol{\epsilon}^r_{it})$ is given by (9). Otherwise simulation proceeds in the same way as without classification error.[19]

5.1 Classification error and the new simulator

For the new simulator, classification error poses a more serious challenge. The reason is that in computing the simulator, *which* option the observed payoff corresponds to influences the distribution from which $\boldsymbol{\epsilon}^r_{it}$ is drawn. But the choice corresponding to the observed payoff

[19] Other forms of classification error will not in general have analytical solutions for the choice probabilities. In that case, they need to be simulated separately.

is a random variable from the point of view of the econometrician when there is classification error in the data.[20] To solve this problem, one can introduce the true choice $\boldsymbol{d}_{it}$ as a conditioning variable prior to drawing the structural disturbances.

By analogous argument to section 3,

$$f_{\boldsymbol{D}^*,V^*(\boldsymbol{d})|\boldsymbol{D}^*}(\boldsymbol{d}_{it}^*, w_{it}^*|\boldsymbol{d}_{i,t-1}^*) = \int \sum_{\boldsymbol{d}_{it}\in D_o} f_{\boldsymbol{D}^*,V^*(\boldsymbol{d}),\boldsymbol{U},\boldsymbol{D}|\boldsymbol{D}^*}(\boldsymbol{d}_{it}^*, w_{it}^*, \boldsymbol{u}_{it}, \boldsymbol{d}_{it}|\boldsymbol{d}_{i,t-1}^*)d\boldsymbol{u}_{it}. \tag{42}$$

Repeated conditioning then yields

$$f_{\boldsymbol{D}^*,V^*(\boldsymbol{d})|\boldsymbol{D}^*}(\boldsymbol{d}_{it}^*, w_{it}^*|\boldsymbol{d}_{i,t-1}^*) = \int \sum_{\boldsymbol{d}_{it}\in D_o} f_{V^*(\boldsymbol{d})}(w_{it}^*)P(\boldsymbol{D}_{it}^* = \boldsymbol{d}_{it}^*|\boldsymbol{D}_{i,t-1}^* = \boldsymbol{d}_{i,t-1}^*, \boldsymbol{D}_{it} = \boldsymbol{d}_{it})$$
$$P(\boldsymbol{D}_{it} = \boldsymbol{d}_{it}|\boldsymbol{E}_{it} = \boldsymbol{\epsilon}_{it})\phi_{E|V^*(\boldsymbol{d})}(\boldsymbol{\epsilon}_{it}|w_{it}^*)d\boldsymbol{\epsilon}_{it}. \tag{43}$$

$f_{V^*(\boldsymbol{d})}$ is the density of $V_{it}^*(\boldsymbol{d}_{it})$, the element of $\boldsymbol{V}_{it}^*$ corresponding to the given actual choice $\boldsymbol{d}_{it}$. $\phi_{E|V^*(\boldsymbol{d})}$ is the density of the error term conditional on the same $V_{it}^*(\boldsymbol{d}_{it})$.

An unbiased and consistent simulator is therefore

$$\tilde{f}^N_{\boldsymbol{D}^*,V^*(\boldsymbol{d})|\boldsymbol{D}^*}(\boldsymbol{d}_{it}^*, w_{it}^*|\boldsymbol{d}_{i,t-1}^*) = \frac{1}{R}\sum_{r=1}^{R}\sum_{m\in M_o} f_{V_m^*}(w_{it}^*)$$
$$P(\boldsymbol{D}_{it}^* = \boldsymbol{d}_{it}^*|\boldsymbol{D}_{i,t-1}^* = \boldsymbol{d}_{i,t-1}^*, D_{mit} = 1)P(D_{mit} = 1|\boldsymbol{E}_{it} = \boldsymbol{\epsilon}_{it}^{rm}). \tag{44}$$

M_o is the set of indices m such that $\omega_m = 1$, i.e. which correspond to the elements of D_o. $P(\boldsymbol{D}_{it}^* = \boldsymbol{d}_{it}^*|\boldsymbol{D}_{i,t-1}^* = \boldsymbol{d}_{i,t-1}^*, D_{mit} = 1)$ is given by (40) and

$$P(D_{mit} = 1|\boldsymbol{E}_{it} = \boldsymbol{\epsilon}_{it}^{rm}) = I(v_{mit}^r > v_{jit}^r \forall\, j \neq m). \tag{45}$$

Each $\boldsymbol{\epsilon}_{it}^{rm}$ is drawn from the distribution with the density $\phi_{E|V_m^*}$, following the procedure described in section 3.3.

While this simulator is unbiased and consistent, it is not particularly efficient. The reason is that the same number of draws are used to evaluate the likelihood contribution of each possible choice, even though it is usually very likely that the actual choice is the observed choice. A more efficient simulator would devote different numbers of draws to each possible choice depending on which choice was observed.

[20] A similar problem arises for the Wolpin Simulator, as the composition of the vector of simulated shocks $\tilde{\epsilon}_{it}^r$ depends on the true choice.

This can again be accomplished using an *importance sampling* approach. Instead of devoting the same number of draws to each possible choice, one can draw a random $m_r \in M_o$ for each r from the distribution with the probability mass function

$$f(m) = \begin{cases} \frac{P(\boldsymbol{D}_{it}^*=\boldsymbol{d}_{it}^*|\boldsymbol{D}_{i,t-1}^*=\boldsymbol{d}_{i,t-1}^*,D_{m_r it}=1)f_{V_{m_r}^*}(w_{it}^*)}{\sum_{m\in M_o} P(\boldsymbol{D}_{it}^*=\boldsymbol{d}_{it}^*|\boldsymbol{D}_{i,t-1}^*=\boldsymbol{d}_{i,t-1}^*,D_{mit}=1)f_{V_m^*}(w_{it}^*)} & \text{if } m = m_r \\ 0 & \text{otherwise.} \end{cases} \tag{46}$$

Then

$$\hat{f}^N_{\boldsymbol{D}^*,V^*(\boldsymbol{d})|\boldsymbol{D}^*}(\boldsymbol{d}_{it}^*, w_{it}^*|\boldsymbol{d}_{i,t-1}^*) = \frac{\sum_{m\in M_o} P(\boldsymbol{D}_{it}^* = \boldsymbol{d}_{it}^*|\boldsymbol{D}_{i,t-1}^* = \boldsymbol{d}_{i,t-1}^*, D_{mit} = 1)f_{V_m^*}(w_{it}^*)}{R} \sum_{r=1}^{R} P(D_{m_r it} = 1|\boldsymbol{E}_{it} = \boldsymbol{\epsilon}_{it}^{rm_r}). \tag{47}$$

is also an unbiased and consistent simulator.

This simulator will be most efficient if $F(m)$ is sampled in even intervals instead of by simple random sampling, as that will minimize the sampling error introduced by randomly drawing m_r. Conceptually, the goal is to achieve sample shares of the different elements of M_o that are as close as possible to their relative probability mass while safeguarding unbiasedness. Various methods are available to achieve this end; the Monte Carlo results below all use *Systematic Sampling* (see e.g. Douc, Cappe and Moulines, 2005).

5.2 Classification error *and* unobserved states

A new simulator that can accommodate both classification error and unobserved state variables can be derived using the same ideas as in the case without unobserved state variables. An individual's likelihood contribution is given by

$$f_{\{\boldsymbol{D}_t^*,W_t^*\}_{t=1}^T}(\{\boldsymbol{d}_{it}^*, w_{it}^*\}_{t=1}^T) = \int \sum_{\{\boldsymbol{d}_{it}\}_{t=1}^T \in D_{1:T}} f_{\boldsymbol{D}_1^*,V_{c1}^*,\boldsymbol{U}_{i1},\boldsymbol{D}_1}(\boldsymbol{d}_{i1}^*, w_{i1}^*, \boldsymbol{u}_{i1}, \boldsymbol{d}_{i1}) \prod_{t=2}^{T} f_{\boldsymbol{D}_t^*,V_{ct}^*,\boldsymbol{U}_t,\boldsymbol{D}_t|\{\boldsymbol{D}_s^*,V_{ct}^*,\boldsymbol{U}_s,\boldsymbol{D}_s\}_{s=1}^{t-1}}(\boldsymbol{d}_t^*, w_{it}^*, \boldsymbol{u}_{it}, \boldsymbol{d}_{it}|\{\boldsymbol{d}_{is}^*, w_{is}^*, \boldsymbol{u}_{is}, \boldsymbol{d}_{is}\}_{s=1}^{t-1}) d\boldsymbol{u}_i. \tag{48}$$

$\{\boldsymbol{d}_{it}\}_{t=1}^{T}$ is a choice sequence from the set of all possible choices histories $D_{1:T}$ given the observations. Appropriate conditioning yields

$$f_{\{\boldsymbol{D}_t^*, W_t^*\}_{t=1}^T}(\{\boldsymbol{d}_{it}^*, w_{it}^*\}_{t=1}^T) = \int \sum_{\{m_t\}_{t=1}^T \in M_{1:T}} f_{V_{m1}^*}(w_{i1}^*) P(\boldsymbol{D}_{i1}^* = \boldsymbol{d}_{i1}^* | D_{mi1} = 1)$$
$$P(D_{mi1} = 1 | \boldsymbol{U}_{i1} = \boldsymbol{u}_{i1}) \phi_{\boldsymbol{U}_1 | V_{m1}^*}(\boldsymbol{u}_{i1} | w_{i1}^*) \prod_{t=2}^{T} f_{V_{mt}^* | \{\boldsymbol{U}_s\}_{s=1}^{t-1}}(w_{it}^* | \{\boldsymbol{u}_{is}\}_{s=1}^{t-1})$$
$$P(\boldsymbol{D}_{it}^* = \boldsymbol{d}_{it}^* | \boldsymbol{D}_{i,t-1}^* = \boldsymbol{d}_{i,t-1}^*, D_{mit} = 1) P(D_{mit} = 1 | \{\boldsymbol{U}_{is} = \boldsymbol{u}_{is}\}_{s=1}^{t})$$
$$\phi_{\boldsymbol{U}_t | V_{mt}^*, \{\boldsymbol{U}_{is}\}_{s=1}^{t-1}}(\boldsymbol{u}_{it} | w_{it}^*, \{\boldsymbol{u}_{is}\}_{s=1}^{t-1}) d\boldsymbol{u}_i. \quad (49)$$

$\{m_t\}_{t=1}^{T}$ is a sequence of indices and $M_{1:T}$ is the set of index sequences corresponding to the possible choice sequences $D_{1:T}$. Hence an unbiased and consistent simulator is

$$\tilde{f}^N(\{\boldsymbol{d}_{it}^*\}_{t=1}^T, \{w_{it}^*\}_{t=1}^T) = \frac{1}{R} \sum_{r=1}^{R} \sum_{\{m_t\}_{t=1}^T \in M_{1:T}} f_{V_{m1}^*}(w_{i1}^*) P(\boldsymbol{D}_{i1}^* = \boldsymbol{d}_{i1}^* | D_{mi1} = 1)$$
$$P(D_{mi1} = 1 | \boldsymbol{U}_{i1} = \boldsymbol{u}_{i1}^{mr}) \prod_{t=2}^{T} f_{V_{mt}^* | \{\boldsymbol{U}_s\}_{s=1}^{t-1}}(w_{it}^* | \{\boldsymbol{u}_{is}^{mr}\}_{s=1}^{t-1})$$
$$P(\boldsymbol{D}_{it}^* = \boldsymbol{d}_{it}^* | \boldsymbol{D}_{i,t-1}^* = \boldsymbol{d}_{i,t-1}^*, D_{mit} = 1) P(D_{mit} = 1 | \{\boldsymbol{U}_{is} = \boldsymbol{u}_{is}^{mr}\}_{s=1}^{t}). \quad (50)$$

where each $\boldsymbol{u}_{it}^{mr}$ is drawn from the distribution with the pdf $\phi_{\boldsymbol{U}_t | V_{mt}^*}(\boldsymbol{u}_{it} | w_{it}^*)$ according to the procedure developed in section 3.3.

This simulator is usually computationally infeasible, as the number of possible choice histories rises exponentially with T. However, an alternative simulator along the lines of (47) is straightforward to construct. Draw $\{m_{rt}\}_{t=1}^{T} \in M_{1:T}$ for each r from the distribution with the probability mass function

$$f(\{m_t\}_{t=1}^T) = \begin{cases} \frac{P(\boldsymbol{D}_{i1}^* = \boldsymbol{d}_{i1}^* | D_{m_{r1} i1} = 1)}{\sum_{m \in M_1} P(\boldsymbol{D}_{i1}^* = \boldsymbol{d}_{i1}^* | D_{mi1} = 1)} \prod_{t=2}^{T} \frac{P(\boldsymbol{D}_{it}^* = \boldsymbol{d}_{it}^* | \boldsymbol{D}_{i,t-1}^* = \boldsymbol{d}_{i,t-1}^*, D_{m_{rt} it} = 1)}{\sum_{m \in M_t} P(\boldsymbol{D}_{it}^* = \boldsymbol{d}_{it}^* | \boldsymbol{D}_{i,t-1}^* = \boldsymbol{d}_{i,t-1}^*, D_{mit} = 1)} & \text{if } \{m_t\}_{t=1}^T = \{m_{rt}\}_{t=1}^T \\ 0 & \text{otherwise.} \end{cases} \quad (51)$$

Then another unbiased and consistent simulator is

$$\hat{f}^N(\{\boldsymbol{d}_{it}^*\}_{t=1}^T, \{w_{it}^*\}_{t=1}^T) = \frac{\left[\sum_{m\in M_1} P(\boldsymbol{D}_{i1}^* = \boldsymbol{d}_{i1}^* | D_{mi1} = 1)\right] \prod_{t=2}^T \sum_{m\in M_t} P(\boldsymbol{D}_{it}^* = \boldsymbol{d}_{it}^* | \boldsymbol{D}_{i,t-1}^* = \boldsymbol{d}_{i,t-1}^*, D_{mit} = 1)}{R}$$
$$\sum_{r=1}^R f_{V_{m_r 1}^*}(w_{i1}^*) P(D_{m_r i1} = 1 | \boldsymbol{U}_{i1} = \boldsymbol{u}_{i1}^r) \prod_{t=2}^T f_{V_{m_r t}^* | \{\boldsymbol{U}_s\}_{s=1}^{t-1}}(w_{it}^* | \{\boldsymbol{u}_{is}^r\}_{s=1}^{t-1}) P(D_{m_r it} = 1 | \{\boldsymbol{U}_{is} = \boldsymbol{u}_{is}^r\}_{s=1}^t). \tag{52}$$

where again each $\boldsymbol{u}_{it}^r$ is drawn from the distribution with the pdf $\phi_{\boldsymbol{U}_t | V_{mt}^*}(\boldsymbol{u}_{it} | w_{it}^*)$ according to the procedure developed in section 3.3.

It is again advisable to add a resampling step when the number of periods is at least moderately large. The particle filter described in section 4.2 can be adapted for this purpose. An algorithm for computing a version of (52) with a resampling step is given in Appendix C.

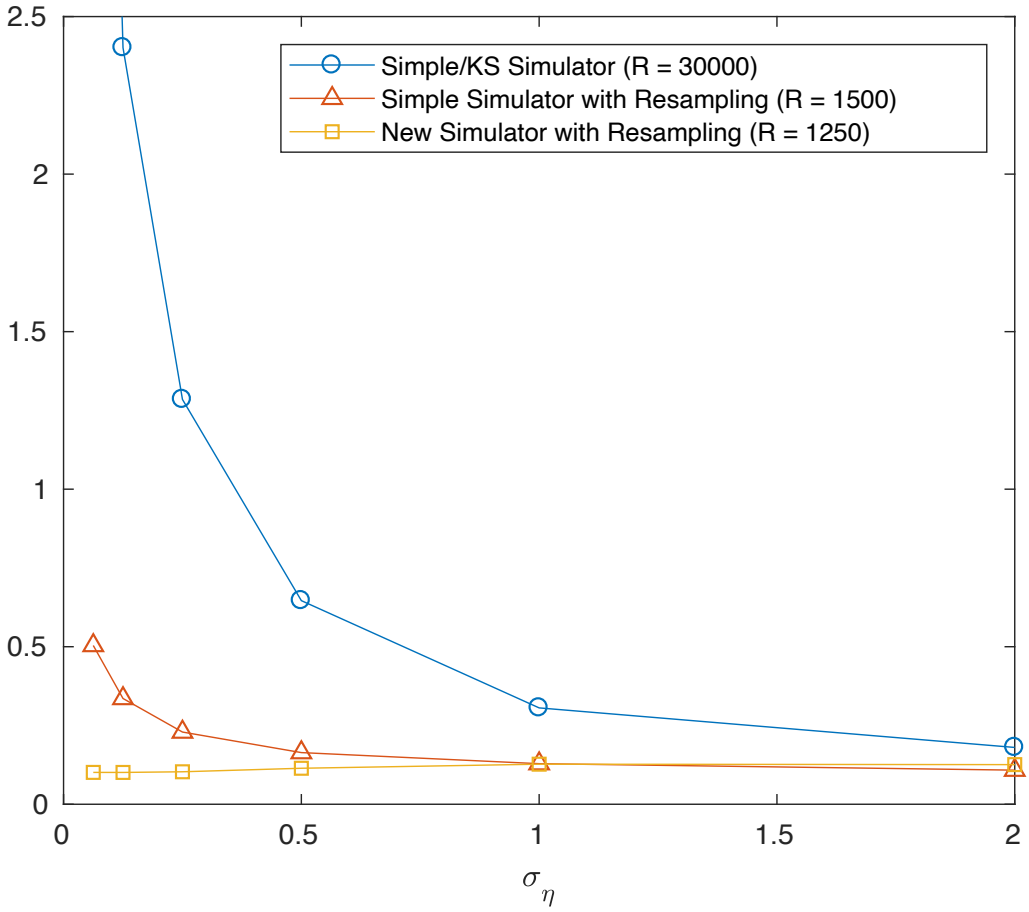
5.3 Monte Carlo evidence: simulation performance

Accounting for classification error in likelihood simulation is more computationally burdensome for the new simulator than for the Simple Simulator. However, even with computation time held constant, the new simulator still performs at least as well as the Simple Simulator with plausible values for the extent of classification error and the standard deviation of the measurement error in the observed payoff.

Figure 5 shows results for the stylized Keane and Wolpin (1994) model with both classification error and unobserved state variables. In contrast to above, I assume that a payoff is observed for all options ($\Omega = I$) to allow the classification error to range across all options. In particular, I assume that the true choice is reported with a probability of $\pi^{corr} = 0.94$ and each alternative choice is reported with a probability of $\pi^{incorr} = 0.02$.[21]

[21] In the terms of (39), this is to say that $\gamma_1 = \log(47)$ and $\gamma_2 = 0$.

Figure 5: Mean standard deviation of $\log(\hat{f})$



Notes. The diagram shows the sample standard deviation of $\log(\hat{f})$ in a model with both unobserved state variables and classification error. All results shown are means for $I = 20$ sequences of $\{\boldsymbol{d}_{it}\}_{t=1}^{T}$ and $\{w_{it}^{*}\}_{t=1}^{T}$, where $T = 5$. In each case, the sample standard deviation is taken over $S = 1000$ evaluations.

The diagram presents results for three simulators. The Simple Simulator without resampling is the method described in Keane and Sauer (2010) and applied in Keane and Wolpin (2001), Sauer (2004), and Sauer (2015). Adding a resampling step leads to large improvements in accuracy even if computation time is held fixed, especially if the standard deviation of the measurement error in the observed payoff σ_η is small. Despite additional computation costs, the new simulator with Resampling leads to further improvements if σ_η is small. The Sequential Monte Carlo version of the new simulator is robust in the sense that it provides acceptable accuracy for all reasonable values of σ_η.[22]

[22]Results for the Wolpin Simulator are not presented, as classification error imposes the same additional computational burden on the Wolpin simulator as on the new simulator.

Table 3: Average computation time for $\hat{f}(\{\boldsymbol{d}_{it}^*\}_{t=1}^T, \{w_{it}^*\}_{t=1}^T)$

	Simple/KS Sim. $R = 30000$	Simple/KS Sim. Res. $R = 1500$	New Sim. Res. $R = 1250$
Time (in ms)	31.232	27.471	26.959

Notes. Average computation time for $\hat{f}(\{\boldsymbol{d}_{it}^*\}_{t=1}^T, \{w_{it}^*\}_{t=1}^T)$ in a model with both unobserved state variables and classification error.

6 Monte Carlo evidence: estimation performance

This section applies the new simulator to an actual estimation problem from the literature and compares its performance to the Simple Simulator and the new simulator. The model used is the full Keane and Wolpin (1994) model with two minor amendments to make it a more suitable test case. It is shown that the new simulator leads to more accurate parameter estimates when the model is estimated on simulated data. This additional accuracy translates into more accurate results in counterfactual simulations in a simple policy experiment.

6.1 Model

The model of Keane and Wolpin (1994) is a model of occupational choice. In each of $T = 40$ discrete periods of time, agents choose between $K = 4$ different options: blue-collar work, white-collar work, education, and home production. The per-period utility functions are

$$u_{1t} = w_{1t} = \exp(\alpha_{10} + \alpha_{11}s_t + \alpha_{12}x_{1t} - \alpha_{13}x_{1t}^2 + \alpha_{14}x_{2t} - \alpha_{15}x_{2t}^2 + \epsilon_{1t}) \tag{53}$$

$$u_{2t} = w_{2t} = \exp(\alpha_{20} + \alpha_{21}s_t + \alpha_{22}x_{2t} - \alpha_{23}x_{2t}^2 + \alpha_{24}x_{1t} - \alpha_{25}x_{1t}^2 + \epsilon_{2t}) \tag{54}$$

$$u_{3t} = \beta_0 - \beta_1 I(s_t \geq 13) - \beta_2(1 - d_{3,t-1}) + \epsilon_{3t} \tag{55}$$

$$u_{4t} = \gamma_0 + \epsilon_{4t} \tag{56}$$

w_{1t} and w_{2t} are the agent's (latent) wages in occupation one and two. s_t is the number of periods of schooling accumulated by the beginning of period t. x_{1t} and x_{2t} are an individual's total periods of work experience at the beginning of period t in occupation one and two, respectively. $\boldsymbol{d}_t$ is a vector of indicator variables, where for each element $d_{kt} = 1$ if option k is chosen in period t and $d_{kt} = 0$ otherwise. Hence $d_{3,t-1}$ is an indicator variable of whether schooling was chosen in the previous period. $\boldsymbol{\epsilon}_t \sim N(0, \Sigma)$ is a vector of serially uncorrelated

shocks, where Σ is parameterized by the parameter vector $\boldsymbol{a}$. $\boldsymbol{\epsilon}_t$ is known to the agent at the beginning of period t but not before. $\boldsymbol{\theta} = \{\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{a}\}$ is the full parameter vector of the model.

Individuals have perfect knowledge of the true model and rational expectations about the future. Let $\boldsymbol{S}_t = \{s_t, x_{1t}, x_{2t}, d_{3,t-1}, \boldsymbol{\epsilon}_t\}$ be the vector of state variables, and denote the vector of pre-determined state variable by $\overline{\boldsymbol{S}}_t = \{s_t, x_{1t}, x_{2t}, d_{3,t-1}\}$. The vector of pre-determined state variables evolves deterministically given the previous period's choices: $x_{1,t+1} = x_{1,t} + d_{1,t}$, $x_{2,t+1} = x_{2,t} + d_{2,t}$, $s_{t+1} = s_t + d_{3,t}$, and $d_{3,t-1}$ is simply the previous period's choice. Initial conditions are $x_{11} = x_{21} = 0$ and $s_1 = 10$. All agents are *ex ante* homogeneous; different choices arise only due to different draws of $\boldsymbol{\epsilon}_t$.

Then the agent's decision problem in period t can be written recursively as

$$\underset{k \leq K}{\operatorname{argmax}}\, V_{kt}(\boldsymbol{S}_t) \tag{57}$$

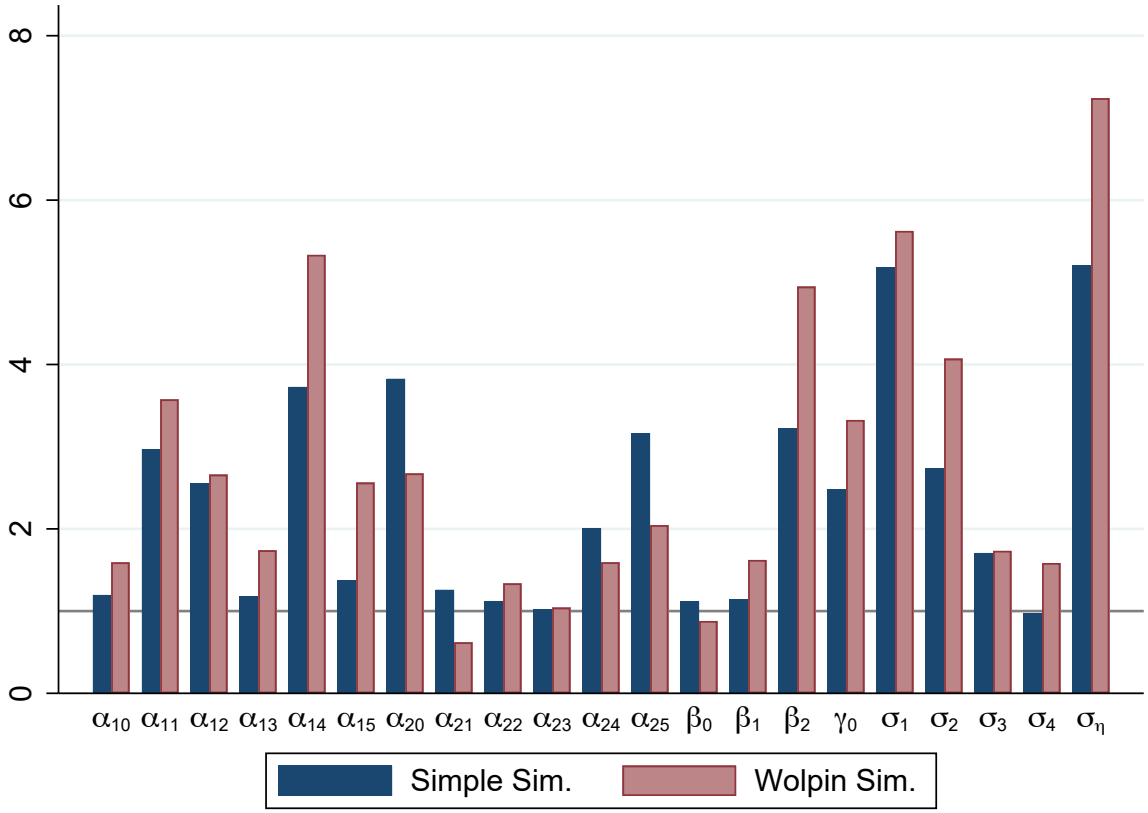
where

$$V_{kt}(\boldsymbol{S}_t) = \begin{cases} u_{kt}(\boldsymbol{S}_t) + \delta \, \mathrm{E}\left[\max_{j \leq K} V_{j,t+1}(\boldsymbol{S}_{t+1}) | \overline{\boldsymbol{S}}_t, d_{kt} = 1\right] & \text{if } t < T \\ u_{kt}(\boldsymbol{S}_t) & \text{if } t = T \end{cases} \tag{58}$$

is the *alternative-specific value function*. E denotes the mathematical expectations operator. The discount factor δ is fixed at $\delta = 0.95$. The agent's problem can be solved by backward recursion using (57) and (58). As there is in general no analytical solution for this expectation in (58), it usually has to be evaluated using numerical integration methods at considerable computational cost.

The version of this model used here departs from the original Keane and Wolpin (1994) paper in two ways. First, as has become standard practice in the subsequent literature starting with Keane and Wolpin (1997), I add a measurement error in wages $\eta \sim N(0, \sigma_\eta^2)$, with the variance σ_η^2 added to θ as an additional model parameter. Second, I assume that Σ is diagonal and not unrestricted as in Keane and Wolpin (1994). This shortens $\boldsymbol{\theta}$ by three parameters and aids identification of the remaining elements.

Figure 6: Relative root mean squared error of parameter estimates



Notes. The diagram shows the relative root mean squared error of the estimates for all model parameters. Root mean squared errors obtained using the new simulator are normalized to one, as indicated by the grey horizontal line. The root mean squared errors are calculated over 20 estimation runs using simulated balanced panels of 10,000 individuals each. In each case, the estimation is started from a different point in the neighbourhood of the true values. For each observation, 100 simple random draws are used to simulate the likelihood.

6.2 Monte Carlo results

Figure 6 shows the relative root mean squared error of the estimates for all model parameters. Root mean squared errors obtained using the new simulator are normalized to one, as indicated by the grey horizontal line. The root mean squared errors are calculated over 20 estimation runs using simulated balanced panels of 10,000 individuals each. In each case, the estimation is started from a different point in the neighbourhood of the true values. For each observation, 100 simple random draws are used to simulate the likelihood.[23]

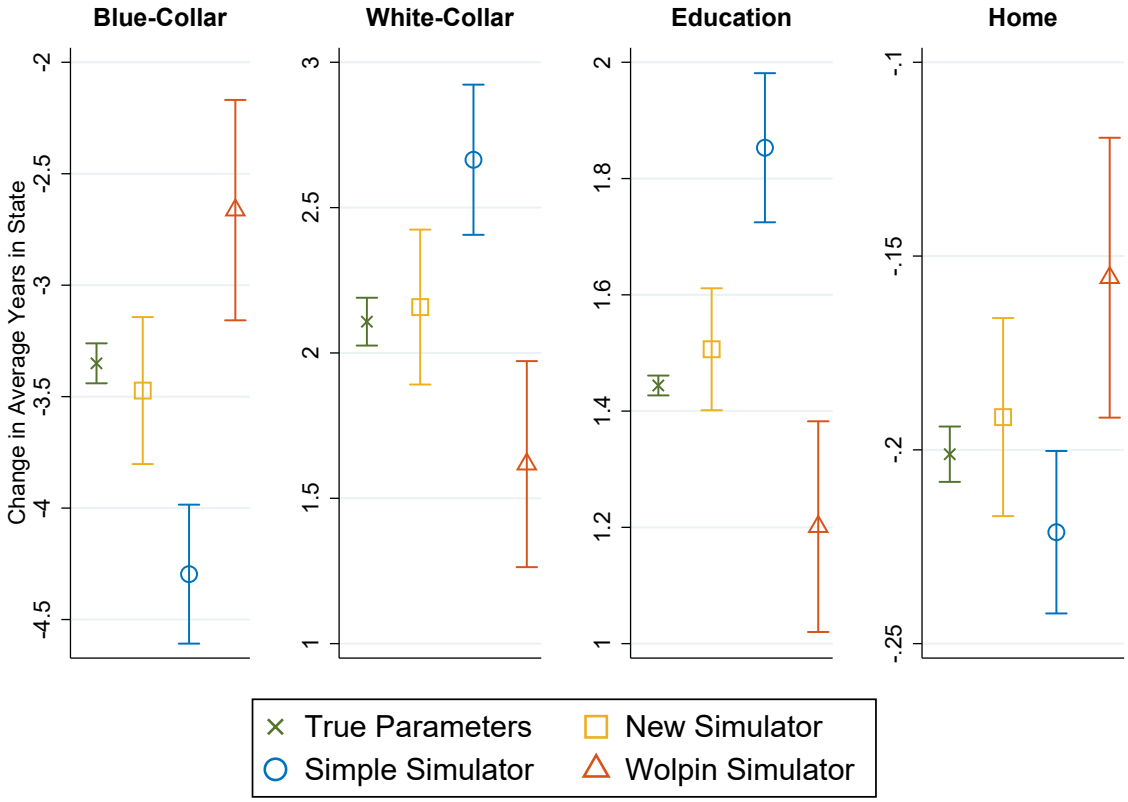
In order to minimize error in the calculation of future values, 1000 Halton draws are used to evaluate each expectation. As is standard in this literature, the Logit kernel smoother proposed by McFadden (1989) is used to smooth the likelihood function, so that a derivative-based

[23]Given the large computational costs of likelihood evaluation, it is not uncommon for applied researchers to use even fewer draws than this. For instance, Cohen-Goldner and Eckstein (2008) use just 25 simple random draws.

optimization algorithm can be used.[24] Numerical maximization of the likelihood is performed using 10 iterations of the BHHH algorithm.

Figure 6 shows that the Simple Simulator and the Wolpin Simulator achieve comparable estimation performance. However, except for a small number of parameters, the root mean squared error of the estimates is much higher than for the new simulator. On average, the root mean squared error of estimated parameters is roughly half as large using the new simulator as with the other methods.

Figure 7: Estimated Impact of a $500 Tuition Subsidy



Notes. Estimated impact of a $500 tuition subsidy on average years spent in each occupation. Markers indicate the mean effect over 20 estimation runs; bars indicate one standard deviation. Estimation uses balanced panels of 10,000 individuals each. In each case, the estimation is started from a different point in the neighbourhood of the true values. For each observation, 100 simple random draws are used to simulate the likelihood. Counterfactual simulations from the model are performed separately for each set of parameters by simulating paths for 10,000 individuals.

This difference in estimation precision can lead to economically meaningful differences in model predictions. Figure 7 shows the impact of a $500 tuition subsidy in the Keane and Wolpin (1994) model. At the true parameters, the subsidy leads to a rise in the average number

[24] Results for a 'GHK-like' smoothing approach that does not introduce bias are presented in Appendix A.

of years individuals spend in education and white-collar work, and a fall in the number of years spent in blue-collar work and home production.

Counterfactual simulations based on the estimated parameters obtained using any of the simulators preserve this qualitative result. However, the new simulator substantially outperforms the other two in terms of bias. On average, a researcher studying the effect of a tuition subsidy on choice shares using the Keane and Wolpin (1994) model would typically have found an effect within 15 percent of the true magnitude if they had estimated the model using the new simulator, but would have been more than 30 percent away from the true magnitude if they had used either of the existing methods.

7 Conclusion

This paper has shown that a new simulator for the likelihood in discrete choice models with observed flow payoffs can lead to large improvements in simulation accuracy. This improved accuracy translates into more efficient and less biased parameter estimates in Simulated Maximum Likelihood estimation. The new simulator can be applied in models with unobserved state variables and classification error, and in conjunction with different smoothing approaches. It is particularly valuable for dynamic models with moderate measurement error in the observed payoffs relative to the standard deviation in the structural disturbances. Moreover, it was shown that a resampling step leads to large improvements in simulation accuracy in models with unobserved state variables when data is observed for even a moderately large number of periods. Resampling can be used in conjunction with any of the likelihood simulators in the literature, including the simulator proposed in this paper.

The results in this paper bolster the case for using Simulated Maximum Likelihood to estimate dynamic discrete choice models when this is feasible, rather than Indirect Inference as has become popular more recently. It was shown that when appropriate simulators are used, an individual's likelihood contribution can often be simulated with high accuracy even in the presence of unobserved state variables and/or classification error. In these cases, Simulated Maximum Likelihood inherits the efficiency properties of Maximum Likelihood and will thus yield more accurate parameter estimates than Indirect Inference.

References

Blevins, Jason R. 2016. "Sequential Monte Carlo methods for estimating dynamic microeconomic models." *Journal of Applied Econometrics*, 31(5): 773–804.

Cohen-Goldner, Sarit, and Zvi Eckstein. 2008. "Labor mobility of immigrants: Training, experience, language, and opportunities." *International Economic Review*, 49(3): 837–872.

Creal, Drew. 2012. "A Survey of Sequential Monte Carlo Methods for Economics and Finance." *Econometric Reviews*, 31(3): 245–296.

Douc, Randal, Olivier Cappe, and Eric Moulines. 2005. "Comparison of resampling schemes for particle filtering." *Proceedings of the 4th International Symposium on Image and Signal Processing and Analysis*, 64–69.

Eckstein, Zvi, and Kenneth I. Wolpin. 1999. "Why Youths Drop out of High School: The Impact of Preferences, Opportunities, and Abilities." *Econometrica*, 67(6): 1295–1339.

Gelman, Andrew, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, and Donald B Rubin. 2013. *Bayesian data analysis.* CRC press.

Gordon, N.J., D.J. Salmond, and A.F.M. Smith. 1993. "Novel approach to nonlinear/non-Gaussian Bayesian state estimation." *IEE Proceedings F (Radar and Signal Processing)*, 140(2): 107.

Hausman, J.A., Jason Abrevaya, and F.M. Scott-Morton. 1998. "Misclassification of the dependent variable in a discrete-response setting." *Journal of Econometrics*, 87(2): 239–269.

Keane, Michael P. 1994. "A Computationally Practical Simulation Estimator for Panel Data." *Econometrica*, 62(1): 95.

Keane, Michael P., and Kenneth I. Wolpin. 1994. "The Solution and Estimation of Discrete Choice Dynamic Programming Models by Simulation and Interpolation: Monte Carlo Evidence." *The Review of Economics and Statistics*, 76(4): 648.

Keane, Michael P., and Kenneth I. Wolpin. 1997. "The Career Decisions of Young Men." *Journal of Political Economy*, 105(3): 473–522.

Keane, Michael P., and Kenneth I. Wolpin. 2001. "The Effect of Parental Transfers and Borrowing Constraints on Educational Attainment." *International Economic Review*, 42(4): 1051–1103.

Keane, Michael P., and Robert M. Sauer. 2010. "a Computationally Practical Simulation Estimation Algorithm for Dynamic Panel Data Models With Unobserved Endogenous." *International Economic Review*, 51(4): 925–958.

Malik, Sheheryar, and Michael K. Pitt. 2011. "Particle filters for continuous likelihood evaluation and maximisation." *Journal of Econometrics*, 165(2): 190–209.

McFadden, Daniel. 1989. "A Method of Simulated Moments for Estimation of Discrete Response Models Without Numerical Integration." *Econometrica*, 57(5): 995.

Sauer, Robert. 1998. "Job Mobility and the Market for Lawyers." *Journal of Political Economy*, 106(1): 147–171.

Sauer, Robert M. 2004. "Educational financing and lifetime earnings." *Review of Economic Studies*, 71(4): 1189–1216.

Sauer, Robert M. 2015. "Does it Pay for Women to Volunteer?" *International Economic Review*, 56(2): 537–564.

Smith, Adrian FM, and Alan E Gelfand. 1992. "Bayesian statistics without tears: a sampling–resampling perspective." *The American Statistician*, 46(2): 84–88.

Stinebrickner, Todd R. 2001. "A Dynamic Model of Teacher Labor Supply." *Journal of Labor Economics*, 19(1): 196–230.

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

Appendix

A GHK-like simulators

A very popular simulator in the wider discrete choice literature is the GHK Simulator (Keane, 1994). The GHK Simulator exploits the structure of Probit models to achieve more efficient simulation of the choice probability. While none of the models discussed in this paper are Probit models and thus amenable to simulation using the traditional GHK algorithm, broadly similar procedures can be used to obtain more efficient probability estimates in more general settings. The key idea is to draw some parts of the vector of errors conditional on information about the observed choice.

As will be shown, 'GHK-like' versions of all three simulators can be constructed that apply this idea. The GHK-like versions are generally more efficient for an equal number of draws, but are also more expensive to compute than the basic versions. An important advantage is that the GHK-like simulators are smooth functions of model parameters, so derivative-based optimization methods can be used without resorting to a smoothing procedure.

A.1 Derivation

The starting point for GHK-like methods is the observation that the event of option c being chosen can be redescribed as the joint occurrence of $N-1$ distinct events, namely for each $j \neq c$, the event that $v_{jit} < v_{cit}$. Define

$$\tilde{d}_{jit} = I(v_{jit} < v_{cit}).$$

Then

$$f_{\boldsymbol{D},V^*(\boldsymbol{d})}(\boldsymbol{d}_{it}, w^*_{it}) = f_{\{\tilde{D}_j\}_{j\neq c},V^*(\boldsymbol{d})}(\{\tilde{d}_{jit}=1\}_{j\neq c}, w^*_{it}) \tag{A.1}$$

and therefore

$$f_{\boldsymbol{D},V^*(\boldsymbol{d})}(\boldsymbol{d}_{it}, w^*_{it}) = \int f_{\{\tilde{D}_j\}_{j\neq c},V^*(\boldsymbol{d}),\boldsymbol{E}}(\{\tilde{d}_{jit}=1\}_{j\neq c}, w^*_{it}, \boldsymbol{\epsilon}_{it}) d\boldsymbol{\epsilon}_{it}. \tag{A.2}$$

Now assume without loss of generality that $c = 1$ (for $c \neq 1$, the covariance matrix Σ can be appropriately re-ordered). Then by repeated conditioning and integrating out the Nth element

of $\boldsymbol{\epsilon}_{it}$, one can obtain

$$f_{\boldsymbol{D},V^*(\boldsymbol{d})}(\boldsymbol{d}_{it}, w^*_{it}) = \int f_{V^*(\boldsymbol{d})|E_1}(w^*_{it}|\epsilon_{1it}) \left[\prod_{n=2}^{N} P(\tilde{D}_{nit} = 1|\{E_{jit} = \epsilon_{jit}\}_{j=1}^{n-1})\right] \phi_{E_1}(\epsilon_{1it}) \prod_{n=2}^{N-1} \phi_{\boldsymbol{E}}(\epsilon_{nit}|\{\epsilon_{jit}\}_{j=1}^{n-1}, \{\tilde{d}_{jit}\}_{j=2}^{n}) d\check{\boldsymbol{\epsilon}}_{it} \quad \text{(A.3)}$$

where $\check{\boldsymbol{\epsilon}}_{it}$ comprises the first $N-1$ elements of $\boldsymbol{\epsilon}_{it}$. Hence a GHK-like version of the simple simulator is

$$\check{f}^{S}_{\boldsymbol{D},V^*(\boldsymbol{d})}(\boldsymbol{d}_{it}, w^*_{it}) = \frac{1}{R}\sum_{r=1}^{R} f_{V^*(\boldsymbol{d})|E_1}(w^*_{it}|\epsilon^{rc}_{1it}) \prod_{n=2}^{N} P(\tilde{D}_{nit} = 1|\{E^c_{jit} = \epsilon^{rc}_{jit}\}_{j=1}^{n-1}) \quad \text{(A.4)}$$

where ϵ^{rc}_{1it} is drawn from its unconditional distribution and each ϵ^{rc}_{jit} for $1 < j < N$ is drawn according to the appropriate conditional distribution by the same procedure as in the GHK algorithm. The index c indicates that these are draws with covariance matrix Σ^c, which is a re-ordered version of Σ such that the first element corresponds to the chosen option.

An analogous procedure yields alternative forms of the Wolpin Simulator and the new simulator. The expression for the Wolpin Simulator

$$\check{f}^{W}_{\boldsymbol{D},V^*(\boldsymbol{d})}(\boldsymbol{d}_{it}, w^*_{it}) = \frac{1}{R}\sum_{r=1}^{R} f_{V^*(\boldsymbol{d})|\tilde{\boldsymbol{E}}}(w^*_{it}|\tilde{\boldsymbol{\epsilon}}^{rc}_{it}) \prod_{n=2}^{N} P\left(\tilde{D}_{nit} = 1|\{\tilde{E}^c_{jit} = \tilde{\epsilon}^{rc}_{jit}\}_{j=1}^{n-1}, w^*_{it}\right) \quad \text{(A.5)}$$

is analogous to the non-GHK case. Similarly, the expression for the new simulator is

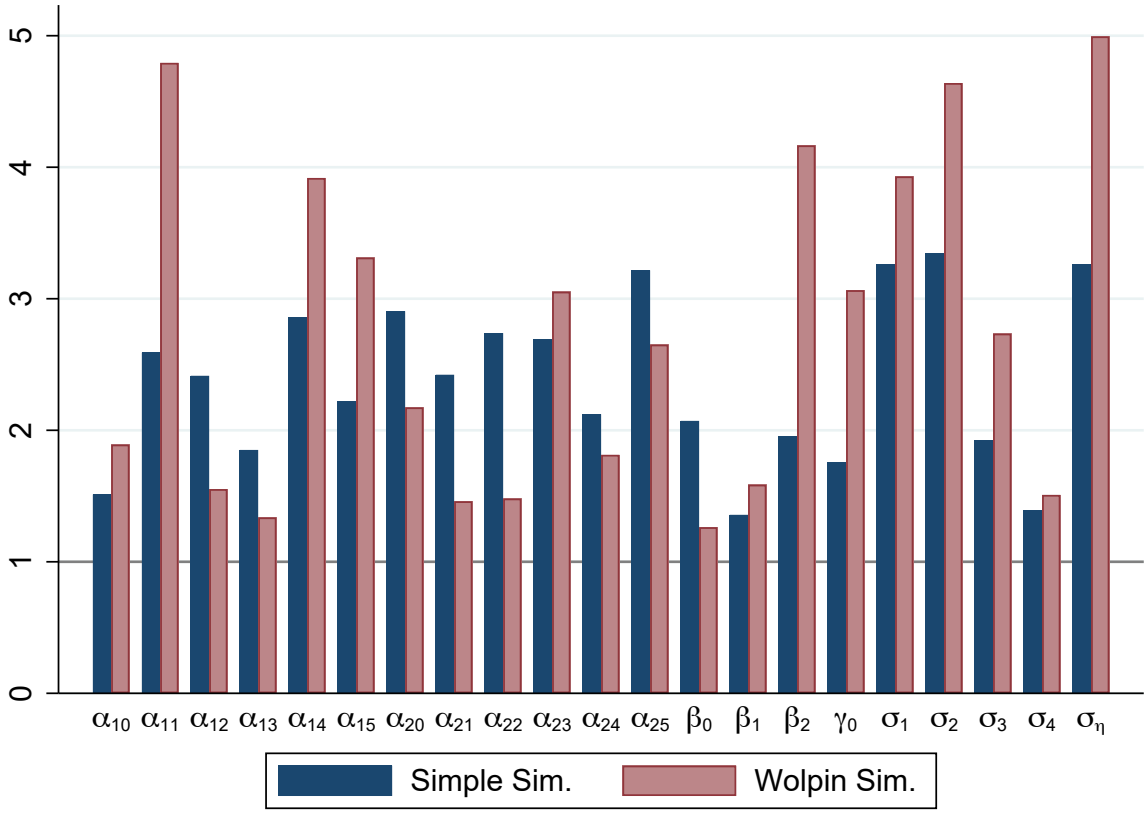
$$\check{f}^{N}_{\boldsymbol{D},V^*(\boldsymbol{d})}(\boldsymbol{d}_{it}, w^*_{it}) = \frac{1}{R}\sum_{r=1}^{R} f_{V^*(\boldsymbol{d})}(w^*_{it}) \prod_{n=2}^{N} P\left(\tilde{D}_{nit} = 1|\{E^c_{jit} = \epsilon^{rc}_{jit}\}_{j=1}^{n-1}\right) \quad \text{(A.6)}$$

where the ϵ^{rc}_{1it} is drawn from its distribution conditional on w^*_{it} using the procedure from section 3.3.

A.2 Monte Carlo estimation results

Monte Carlo estimation results from the Keane and Wolpin (1994) model are very similar using a GHK-like simulator to the results using a Logit smoother that were presented in the main text. The only substantial difference is a somewhat improved accuracy of the new simulator. This is intuitive: as the new simulator offers greater simulation accuracy, the accuracy losses introduced by the Logit smoother become relatively more important.

Figure 8: GHK-like Simulator: Relative root mean squared error of parameter estimates

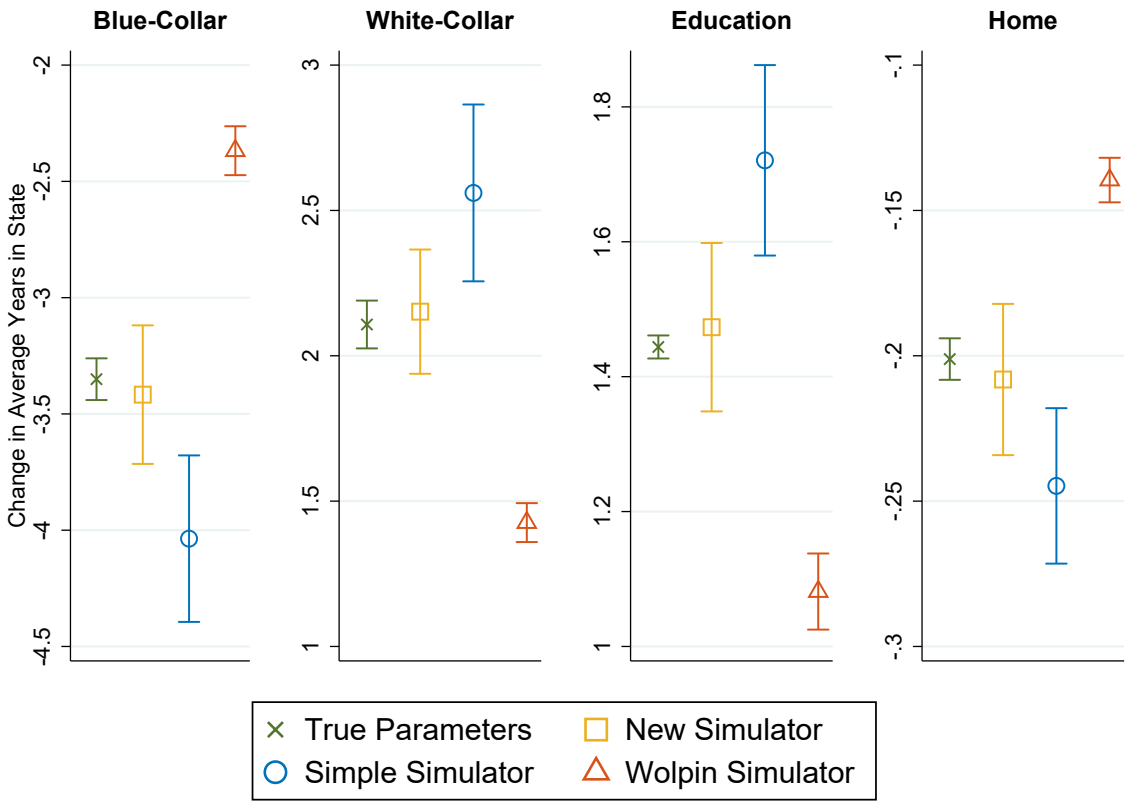


Notes. The diagram shows the relative relative root mean squared error for the estimates for all model parameters. Root mean squared errors obtained using the new simulator are normalized to one, as indicated by the grey horizontal line. The root mean squared errors are calculated over 20 estimation runs using simulated balanced panels of 10,000 individuals each. In each case, the estimation is started from a different point in the neighbourhood of the true values. For each observation, 100 simple random draws are used to simulate the likelihood.

Figure 8 is the analogue of Figure 6 in the main text. It shows the root mean squared errors of parameter estimates relative to the new simulator. In the GHK-like version, the new simulator achieves a lower root mean squared error for all parameters.

Figure 9 is the analogue of Figure 7 in the main text. It shows the estimated impact of a \$500 tuition fee in the Keane and Wolpin (1994) model. Average effects estimated using the GHK-like version of the new simulator are even closer to simulated effects using the true parameters than if a Logit smoother is used.

Figure 9: GHK-like Simulator: Estimated Impact of a $500 Tuition Subsidy



Notes. Estimated impact of a $500 tuition subsidy on average years spent in each occupation. Markers indicate the mean effect over 20 estimation runs; bars indicate one standard deviation. Estimation uses balanced panels of 10,000 individuals each. In each case, the estimation is started from a different point in the neighbourhood of the true values. For each observation, 100 simple random draws are used to simulate the likelihood. Counterfactual simulations from the model are performed separately for each set of parameters by simulating paths for 10,000 individuals.

B Resampling algorithm for unobserved state variables

Define

$$U^*_{cit} = \Phi_c^{-1}(V^*_{it}(\boldsymbol{d}_{it}); \boldsymbol{s}_{it}) - \rho E_{ci,t-1}. \tag{B.1}$$

Conditional on $E_{ci,t-1}$, U^*_{cit} is distributed $N(0, \sigma_c^2 + \sigma_\eta^2)$. Then a computationally convenient algorithm proceeds as follows:

1. Draw R vectors of structural disturbances $\boldsymbol{\epsilon}^r_{i1} = \boldsymbol{u}^r_{i1}$ from the distribution with the pdf $\phi_{\boldsymbol{U}_1|V^*_1(\boldsymbol{d}_1)}(\boldsymbol{u}_{i1}|w^*_{i1})$.

2. For each r, evaluate

$$\tilde{\omega}_{i1}^r = f_{V_1^*(\boldsymbol{d}_1)}(w_{i1}^*)P(\boldsymbol{D}_{i1} = \boldsymbol{d}_{i1}|\boldsymbol{E}_{i1} = \boldsymbol{\epsilon}_{i1}^r).$$

For any subsequent period $t = 2, \ldots, T$, repeat steps 3 to 7:

3. *Optional resampling:* Draw R times from the distribution characterized by the probability mass function

$$f(\boldsymbol{\epsilon}_{i,t-1}) = \begin{cases} \frac{\tilde{\omega}_{i,t-1}^r}{\sum_{r=1}^R \tilde{\omega}_{i,t-1}^r} & \text{if } \boldsymbol{\epsilon}_{i,t-1} = \boldsymbol{\epsilon}_{i,t-1}^r \\ 0 & \text{otherwise.} \end{cases}$$

   A non-random sampling procedure such as Systematic Sampling should be used to minimize the sampling variance. Replace the existing $\{\boldsymbol{\epsilon}_{i,t-1}^r\}_{r=1}^R$ by the sample and replace all $\tilde{\omega}_{i,t-1}^r$ by $\frac{1}{R}\sum_{s=1}^R \tilde{\omega}_{i,t-1}^s$.

4. If a wage is observed, calculate

$$u_{cit}^{r*} = \Phi_c^{-1}(w_{it}^*; \boldsymbol{s}_{it}) - \rho\epsilon_{ci,t-1}^r$$

   for each r. u_{cit}^{r*} summarizes all of the information on $\boldsymbol{u}_{it}^r$ contained in the observed wage and past simulated structural shocks.

5. If a wage is observed, draw a $\boldsymbol{u}_{it}^r$ from

$$\phi_{\boldsymbol{U}_t|V_t^*(\boldsymbol{d}_t),\{\boldsymbol{U}_{is}\}_{s=1}^{t-1}}(\boldsymbol{u}_{it}|w_{it}^*, \{\boldsymbol{u}_{is}\}_{s=1}^{t-1}) = \phi_{\boldsymbol{U}_t|U_c^*}(\boldsymbol{u}_{it}|u_{cit}^{r*})$$

   for each r using the procedure outlined in section 3.3. Otherwise, draw from the unconditional distribution of $\boldsymbol{u}_{it}$.

6. Calculate

$$\boldsymbol{\epsilon}_{it}^r = \rho\boldsymbol{\epsilon}_{i,t-1}^r + \boldsymbol{u}_{it}^r$$

   for each r.

7. Evaluate

$$\begin{aligned}\tilde{\omega}_{it}^r &= f_{V^*(\boldsymbol{d})|\{\boldsymbol{U}_s\}_{s=1}^{t-1}}(w_{it}^*|\{\boldsymbol{u}_{is}^r\}_{s=1}^{t-1})P(\boldsymbol{D}_{it}=\boldsymbol{d}_{it}|\{\boldsymbol{U}_{is}=\boldsymbol{u}_{is}^r\}_{s=1}^{t})\tilde{\omega}_{i,t-1}^r \\ &= f_{U_c^*}(u_{cit}^{r*})P(\boldsymbol{D}_{it}=\boldsymbol{d}_{it}|\boldsymbol{E}_{it}^r=\boldsymbol{\epsilon}_{it}^r)\tilde{\omega}_{i,t-1}^r.\end{aligned}$$

for each r if a wage is observed and

$$\tilde{\omega}_{it}^r = P(\boldsymbol{D}_{it}=\boldsymbol{d}_{it}|\boldsymbol{E}_{it}^r=\boldsymbol{\epsilon}_{it}^r)\tilde{\omega}_{i,t-1}^r$$

otherwise.

Then if no resampling was performed,

$$\hat{f}^S_{\{\boldsymbol{D}_t,W_t^*\}_{t=1}^T}(\{\boldsymbol{d}_{it}\}_{t=1}^T,\{w_{it}^*\}_{t=1}^T) = \frac{1}{R}\sum_{r=1}^R \tilde{\omega}_{iT}^r \tag{B.2}$$

gives the simulator in (30). With resampling in any or all periods, (B.2) yields different unbiased and consistent simulators.

C Resampling algorithm for unobserved state variables and classification Error

A practicable algorithm for computing $\hat{f}^N(\{\boldsymbol{d}_{it}^*\}_{t=1}^T,\{w_{it}^*\}_{t=1}^T)$ proceeds as follows.

1. Draw R times from the distribution with the probability mass function $f(\{m_{rt}\}_{t=1}^T)$ given in (51) using *Systematic Sampling*.[25]

2. Draw R vectors of structural disturbances $\boldsymbol{\epsilon}_{i1}^r = \boldsymbol{u}_{i1}^r$ from the distribution with the pdf $\phi_{\boldsymbol{U}_1|V_{m_{r1}}^*}(\boldsymbol{u}_{it}|w_{i1}^*)$.

3. For each r, evaluate

$$\tilde{\omega}_{i1}^r = f_{V_{m_{rt}1}^*}(w_{i1}^*)P(D_{m_{rt}i1}=1|\boldsymbol{U}_{i1}=\boldsymbol{u}_{i1}^r)$$

[25] If this is infeasible because $M_{1:T}$ is large, an alternative is to draw R samples from the distribution with the probability mass function

$$f(m_{rt}) = \frac{P(\boldsymbol{D}_{it}^*=\boldsymbol{d}_{it}^*|\boldsymbol{D}_{i,t-1}^*=\boldsymbol{d}_{i,t-1}^*,D_{m_{rt}it}=1)}{\sum_{m\in M_t}P(\boldsymbol{D}_{it}^*=\boldsymbol{d}_{it}^*|\boldsymbol{D}_{i,t-1}^*=\boldsymbol{d}_{i,t-1}^*,D_{mit}=1)}$$

for each time period t, ensuring that the draws are uncorrelated over time, e.g. by scrambling the order of the draws. Then each sequence $\{m_{rt}\}_{t=1}^T$ will be distributed according to (51).

where $P(D_{m_r i1} = 1|\boldsymbol{U}_{i1} = \boldsymbol{u}_{i1}^r)$ is calculated as in (45).

Then for any subsequent period $t = 2, \ldots, T$, repeat steps 4 to 8:

4. *Optional resampling:* Draw R times from the distribution characterized by the probability mass function

$$f(\boldsymbol{\epsilon}_{i,t-1}) = \begin{cases} \frac{\tilde{\omega}_{i,t-1}^r}{\sum_{r=1}^R \tilde{\omega}_{i,t-1}^r} & \text{if } \boldsymbol{\epsilon}_{i,t-1} = \boldsymbol{\epsilon}_{i,t-1}^r \\ 0 & \text{otherwise.} \end{cases}$$

   Again a non-random sampling procedure such as Systematic Sampling should be used to minimize the sampling variance. Replace the existing $\{\boldsymbol{\epsilon}_{i,t-1}^r\}_{r=1}^R$ by the sample and replace all $\tilde{\omega}_{i,t-1}^r$ by $\frac{1}{R}\sum_{s=1}^R \tilde{\omega}_{i,t-1}^s$.

5. If a wage is observed in t, calculate

$$u_{it}^{r*} = \Phi_c^{-1}(w_{it}^*; \boldsymbol{s}_{it}) - \rho\epsilon_{m_r i,t-1}^r$$

   for each r. u_{it}^{r*} summarizes all of the information on $\boldsymbol{u}_{it}^r$ contained in the observed wage and past simulated structural shocks.

6. If a wage is observed in t, draw from

$$\phi_{\boldsymbol{U}_t|V_{mt}^*,\{\boldsymbol{U}_{is}\}_{s=1}^{t-1}}(\boldsymbol{u}_{it}|w_{it}^*, \{\boldsymbol{u}_{is}\}_{s=1}^{t-1}) = \phi_{\boldsymbol{U}_t|U_t^*}(\boldsymbol{u}_{it}|u_{it}^{r*})$$

   for each r using the procedure outlined in section 3.3.

7. Calculate

$$\boldsymbol{\epsilon}_{it}^r = \rho\boldsymbol{\epsilon}_{i,t-1}^r + \boldsymbol{u}_{it}^r$$

   for each r.

8. Evaluate

$$\begin{aligned} \tilde{\omega}_{it}^r &= f_{V_{m_r t}^*|\{\boldsymbol{U}_s\}_{s=1}^{t-1}}(w_{it}^*|\{\boldsymbol{u}_{is}^r\}_{s=1}^{t-1}) P(D_{m_r it} = 1|\{\boldsymbol{U}_{is} = \boldsymbol{u}_{is}^r\}_{s=1}^t)\tilde{\omega}_{i,t-1}^r \\ &= f_{U_t^*}(u_{it}^{r*}) P(D_{m_r it} = 1|\boldsymbol{E}_{it}^r = \boldsymbol{\epsilon}_{it}^r)\tilde{\omega}_{i,t-1}^r. \end{aligned}$$

for each r if a wage is observed in t and

$$\tilde{\omega}_{it}^r = P(D_{m_r it} = 1 | \boldsymbol{E}_{it}^r = \boldsymbol{\epsilon}_{it}^r) \tilde{\omega}_{i,t-1}^r.$$

otherwise.

Then if no resampling was performed,

$$\tilde{f}(\{\boldsymbol{d}_{it}^*\}_{t=1}^T, \{w_{it}^*\}_{t=1}^T) = \frac{\left[\sum_{m \in M_1} P(\boldsymbol{D}_{i1}^* = \boldsymbol{d}_{i1}^* | D_{mi1} = 1)\right] \prod_{t=2}^T \sum_{m \in M_t} P(\boldsymbol{D}_{it}^* = \boldsymbol{d}_{it}^* | \boldsymbol{D}_{i,t-1}^* = \boldsymbol{d}_{i,t-1}^*, D_{mit} = 1)}{R} \sum_{r=1}^R \tilde{\omega}_{iT}^r. \tag{C.1}$$

gives the same simulator as (52). With resampling in any or all periods, (C.1) yields different unbiased and consistent simulators.

D Monte Carlo results: alternative model specification

The alternative specification is identical to the basic specification from section 3.4, except that

$$\Phi_n(E_{nit}) = \begin{cases} \exp(E_{nit}) & \text{for } n = 1 \\ \exp(E_{nit}/2) & \text{for } n = 2 \\ 1 + E_{nit} & \text{for } n = 3 \\ 3/2 + E_{nit} & \text{for } n = 4, \end{cases} \tag{D.1}$$

and

$$\Sigma = \begin{pmatrix} 1 & -1/2 & 1/4 & -1/8 \\ -1/2 & 1 & -1/2 & 1/4 \\ 1/4 & -1/2 & 1 & -1/2 \\ -1/8 & 1/4 & -1/2 & 1 \end{pmatrix}.$$

This specification is constructed to have fewer of the special features that mark out the basic specification as unusual.

Monte Carlo results are similar, as shown in Table 4. All simulators take somewhat longer to compute for a general covariance matrix of the structural shocks. Performance of the Wolpin

Table 4: Mean standard deviation of $\log(\hat{f})$, alternative specification

	Simple Sim. $R = 500$	Wolpin Sim. $R = 500$	New Sim. $R = 500$	New Sim. $R = 1000$
$\sigma_\eta = 1/16$	0.329	0.066	0.047	0.033
$\sigma_\eta = 1/8$	0.209	0.066	0.047	0.033
$\sigma_\eta = 1/4$	0.142	0.068	0.048	0.033
$\sigma_\eta = 1/2$	0.101	0.077	0.051	0.036
$\sigma_\eta = 1$	0.080	0.096	0.055	0.040
$\sigma_\eta = 2$	0.073	0.136	0.060	0.044
$\sigma_\eta = 4$	0.069	0.200	0.064	0.046
Time (in ms)	0.253	0.347	0.232	0.262

Notes. All results shown are means for $I = 20$ pairs of $\boldsymbol{d}_{it}$ and w_{it}^*. In each case, the sample standard deviation is taken over $S = 5000$ evaluations.

Simulator deteriorates significantly: It takes longer to compute and is substantially less accurate for all values of σ_η. As a result, and in contrast to the case when structural shocks are uncorrelated, the Wolpin Simulator does not match the performance of the new simulator with the same number of draws even as $\sigma_\eta \to 0$. Intuitively, the reason is that the new simulator takes the information in w_{it}^* on all elements of $\boldsymbol{\epsilon}_{it}$ into account in sampling, whereas the Wolpin Simulator only takes it into account for the element of $\boldsymbol{\epsilon}_{it}$ corresponding to the chosen option.

II. Generalized indirect inference in dynamic structural models: an application to the long-run effect of a conditional cash transfer scheme*

JACK BRITTON
IFS

BEN WALTMANN
Oxford and IFS

Abstract

The Generalized Indirect Inference (GII) method of Bruins et al. (2018) substantially eases the computational burden of estimating discrete choice models using Indirect Inference. However, it has so far been impossible to apply the GII method to an important class of structural dynamic discrete choice models that feature both unobserved heterogeneity and payoffs associated with different options that can depend on a long history of past choices. This paper shows how the GII method can be adapted to make the estimation of such models possible. We compare our method to Simulated Maximum Likelihood and conventional GII in a structural dynamic discrete choice setting where all three methods are feasible. Our new method roughly matches the other two in terms of both estimation accuracy and performance in a simple policy experiment. We then use the method to estimate a structural model of a conditional cash transfer scheme in the UK, which could not be estimated using either of the other methods. We find that the programme, which paid young people up to £3,600 to stay in education or training beyond age 16, increased attendance among students eligible for the full grant by up to 2.2 percentage points. While the policy only had a long-run impact for a small minority of students, wages for that group were 9 percent higher at age 26 than in a counterfactual scenario without the grant.

*We would like to thank James Duffy, Hamish Low, and seminar participants at the Institute for Fiscal Studies for helpful comments and suggestions. The use of the University of Oxford Advanced Research Computing (ARC) facility is gratefully acknowledged (DOI code: http://dx.doi.org/10.5281/zenodo.22558).

1 Introduction

In many economic models, traditional estimation approaches such as Maximum Likelihood or the Method of Moments are prohibitively difficult to implement, but simulation from the model at a given parameter vector is straightforward. A common approach to estimation in these settings is Indirect Inference (Gourieroux, Monfort and Renault, 1993; Smith Jr, 1993). Indirect Inference relies on the estimation of an auxiliary descriptive model that captures the relevant patterns in the data. Estimates of the structural parameters are obtained by minimizing an appropriate measure of the distance between the parameters of the auxiliary model estimated on the actual data, and the parameters of the same model when estimated on simulated data from the structural model. A special case is the Simulated Method of Moments (SMM), where the descriptive model consists of a vector of moments.

A major drawback of using Indirect Inference for estimating discrete choice models is that small changes in the structural parameters lead to discrete changes of the simulated data, so the objective function to be minimized becomes a discontinuous function of the structural parameters. As a result, gradient-based optimization methods cannot be used and researchers need to resort to much slower derivative-free methods, limiting the complexity of models that can be estimated. Furthermore, a non-smooth objective function can make it more difficult to verify convergence and calculate standard errors. Despite these drawbacks, Indirect Inference is widely used to estimate discrete choice models; recent examples include Li and Zhang (2015), Eisenhauer, Heckman and Mosso (2015), Skira (2015), Blundell et al. (2016), and Blundell et al. (2021).

Generalized Indirect Inference (GII) is a variation of this procedure that creates a smooth objective surface. The auxiliary model is estimated on a smoothed version of the simulated data, whereby any binary outcome variable is replaced by a continuous value on the unit interval based on information about latent variables. Smoothing is done in such a way that the vector of estimated auxiliary parameters is similar to (and asymptotically the same as) the parameter vector that would be obtained without smoothing. As shown in Bruins et al. (2018), under certain conditions this GII estimator attains the same limiting distribution as the standard indirect inference estimator. The GII method has been widely used in empirical work; recent important papers include Altonji, Smith Jr and Vidangos (2013), Ulyssea (2018), Morten (2019) and Abaluck and Adams-Prassl (2021).

However, the conventional GII method has important limitations. If the payoffs associated with different options in some periods depend on a long history of past discrete choices — as is common in structural dynamic discrete choice models in labour economics — the conventional version of the GII estimator can suffer from serious finite sample biases, as noted by Altonji, Smith Jr and Vidangos (2013) and Bruins et al. (2018). Moreover, in models with such history dependence where agents are also forward-looking, solving the model can become prohibitively difficult, as smoothing leads to a continuous state space. In models without unobserved heterogeneity between agents, these problems can be overcome by using *one-step-ahead* simulations, i.e. by using observed past choices rather than simulated past choices in simulation; but as most models in applied work do feature unobserved heterogeneity, the usefulness of this approach is limited.

In this paper, we develop a computationally practicable variation of the GII method for use in contexts that exhibit history dependence *and* unobserved heterogeneity. We compare our method to Simulated Maximum Likelihood and conventional GII using *one-step-ahead* simulations in the estimation of the model of Keane and Wolpin (1994), a canonical structural dynamic discrete choice model that can be estimated by all three methods. Our new method roughly matches the other two in terms of both estimation accuracy and performance in a simple policy experiment.

We then employ the new method to estimate a structural model of the career choices of young people in the UK (building on Eckstein and Wolpin, 1999), which would be prohibitively difficult to estimate using conventional GII. We use this model to assess the effects of the Education Maintenance Allowance (EMA), a conditional cash transfer (CCT) scheme that paid young people up to £3,600 ($4,800) for staying in education or training between the ages of 16 and 19. A structural approach allows us to assess the long-run impact of the programme and enables us to identify the main mechanisms of action.

We find that the EMA modestly increased participation in education or training among eligible students, with gains concentrated among those eligible for the full grant of £30 per week (£1200 per year). Among the latter group, which makes up 36 percent of all individuals in the cohort we study, the programme had the largest impact at age 17/18, increasing attendance by 2.2 percentage points. These findings are consistent with previous results from the pilot phase of the EMA (Dearden et al., 2009) and from similar schemes in Australia (Dearden and Heath, 1996) and New York City (Riccio et al., 2013). They complement a large volume of evidence

showing that CCTs are effective for raising school enrolment in developing countries.[1]

Our main focus, however, is the long-run effect of the EMA on labour market outcomes. The presumed effects of conditional cash transfers on long-run outcomes are crucial for the justification of these schemes, and yet they have rarely been studied.[2] We find that among the group of less than 2 percent of young people that were affected by the programme in the long term, employment at age 26 was nearly unchanged, but earnings were on average 9 percent higher than in a counterfactual scenario without the grant, even though additional time in education on average displaced a roughly equal amount of time accumulating work experience. An important factor explaining this result is that most individuals in this group acquired a disproportional amount of work experience even with the EMA, and we estimate rapidly diminishing returns to work experience. However, estimated effects are heterogeneous, with nearly half of individuals in this group earning less at age 26 as a result of the programme. Lower earnings at age 26 as a result of the programme are largely explained by individuals failing to successfully complete the additional years of education for which they enrolled.

The remainder of the paper is laid out as follows. Section 2 briefly recapitulates the Indirect Inference and GII methods before introducing our extended GII method. Section 3 presents evidence that the new method achieves similar performance to more traditional methods in a canonical model of career choice. In section 4, we use the new method to estimate a structural model of young people's schooling and labour market decisions in the UK and investigate the effect of the Education Maintenance Allowance. Section 5 concludes.

2 GII for structural dynamic discrete choice models

In this section, we illustrate our estimation method using a generic model as an example. We use this model to briefly recapitulate Indirect Inference (Gourieroux, Monfort and Renault, 1993) and Generalized Indirect Inference (Bruins et al., 2018), before moving on to our own method.[3]

Consider a model whose outputs are the discrete observables $\boldsymbol{d}_{it}$. Suppose that $\boldsymbol{d}_{it}$ describes one set of mutually exclusive alternatives such that $d_{itk} = 1$ if and only if alternative k is

[1] See Schultz (2004), De Brauw and Hoddinott (2011) and Behrman et al. (2012) for Mexico, Attanasio et al. (2005) for Colombia, Glewwe and Kassouf (2012) for Brazil, and Maluccio and Flores (2005) for Nicaragua.

[2] An exception is Behrman, Parker and Todd (2011), who examine the labour market impacts of the PROGRESA programme in Mexico. However, their analysis is limited by the fact that they cannot look beyond programme participants' late teens.

[3] The exposition of Indirect Inference and GII largely follows Bruins et al. (2018).

observed.[4] Suppose $\boldsymbol{d}_{it}$ is determined by the vector of corresponding latent variables $\boldsymbol{u}_{it}$ such that

$$d_{itk} = 1(u_{itk} > \max_{j \neq k} u_{itj}) = \begin{cases} 1 & \text{if } u_{itk} > \max_{j \neq k} u_{itj} \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

In a discrete choice context, $\boldsymbol{d}_{it}$ could be a set of mutually exclusive options, and $\boldsymbol{u}_{it}$ their associated utilities. Suppose $\boldsymbol{u}_{it} = f(\boldsymbol{x}_{it}, \boldsymbol{\epsilon}_{it}, \mathfrak{d}_{i,t-1}; \boldsymbol{\theta})$, where $\boldsymbol{x}_{it}$ is a vector of observed exogenous variables, $\boldsymbol{\epsilon}_{it}$ is a vector of random shocks, $\mathfrak{d}_{i,t-1} = \{\boldsymbol{d}_{i,s}\}_{s=1}^{t-1}$ is the sequence of past discrete outcomes, and $\boldsymbol{\theta}$ is the vector of structural parameters to be estimated.

2.1 Indirect Inference

In order to estimate $\boldsymbol{\theta}$ by Indirect Inference, one needs to specify an auxiliary statistical model that is straightforward to estimate and sufficiently rich to capture the relevant patterns in the data. For instance, an auxiliary statistical model for the generic model outlined above might include OLS regressions of the elements of $\boldsymbol{d}_{it}$ on the elements of $\boldsymbol{x}_{it}$ and/or $\mathfrak{d}_{i,t-1}$. Let $\boldsymbol{\varphi}_0$ be the vector of 'pseudo-true' parameters of this auxiliary model.[5]

The idea of the Indirect Inference estimator is to pick the structural parameters $\boldsymbol{\theta}$ in such a way that data simulated from the model at $\boldsymbol{\theta}$ resembles the actual data as closely as possible when seen through the lens of the auxiliary model. That is to say, the Indirect Inference estimator $\hat{\boldsymbol{\theta}}^{II}$ minimizes a metric of distance between the vector of parameter estimates from estimating the auxiliary model on the simulated data $\check{\boldsymbol{\varphi}}(\boldsymbol{\theta})$ and the vector of parameter estimates from estimating the auxiliary model on the actual data $\hat{\boldsymbol{\varphi}}$.

A common choice for this metric, which we adopt in the rest of the paper, is the 'Wald' criterion function[6]

$$C(\boldsymbol{\theta}) = (\check{\boldsymbol{\varphi}}(\boldsymbol{\theta}) - \hat{\boldsymbol{\varphi}})' \mathcal{W} (\check{\boldsymbol{\varphi}}(\boldsymbol{\theta}) - \hat{\boldsymbol{\varphi}}). \tag{2}$$

Different choices are possible for the weight matrix $\mathcal{W}$, which will affect the efficiency and small sample properties of the Indirect Inference estimator. A common choice we adopt in the

[4] We limit this model to a single set of mutually exclusive discrete alternatives for ease of exposition. As will be seen below, the ideas apply *mutatis mutandis* to models with multiple sets of mutually exclusive outcomes and models with a mixture of discrete and continuous outcomes.

[5] Simulated Method of Moments is a particular approach to (Generalized) Indirect Inference where the parameters of the auxiliary model are moments of the data.

[6] Two alternatives are the Likelihood Ratio and Lagrange Multiplier criterion functions. All three approaches are discussed in Bruins et al. (2018).

rest of the paper is the inverse of the diagonal of the sample covariance matrix of $\hat{\boldsymbol{\varphi}}$.[7]

As the sample size grows large, $\hat{\boldsymbol{\varphi}}$ converges in probability to the pseudo-true parameters of the auxiliary model $\boldsymbol{\varphi}_0$. Similarly, $\breve{\boldsymbol{\varphi}}(\boldsymbol{\theta})$ converges in probability to the non-stochastic binding function $\boldsymbol{\varphi}(\boldsymbol{\theta})$. For a sufficiently rich auxiliary model, the minimizer $\hat{\boldsymbol{\theta}}^{II}$ of the criterion function will then converge in probability to the true vector of structural parameters $\boldsymbol{\theta}_0$.

In practice, it is usually advisable to estimate the auxiliary parameters on multiple independent simulated data sets with the same number of observations as the actual data and take an average of the estimates to obtain

$$\breve{\boldsymbol{\varphi}}(\boldsymbol{\theta}) = \frac{1}{M} \sum_{m=1}^{M} \breve{\boldsymbol{\varphi}}^m(\boldsymbol{\theta}). \tag{3}$$

The higher M, the lower the variance of $\breve{\boldsymbol{\varphi}}(\boldsymbol{\theta})$ and thus the variance of $\hat{\boldsymbol{\theta}}^{II}$. Selecting an optimal value of M involves trading off the gain in efficiency with the additional computation time required for simulations.

A drawback of this method is that the criterion function $C(\boldsymbol{\theta})$ is generally not differentiable when there are discrete outcomes. Small changes in $\boldsymbol{\theta}$ lead to discrete changes in the discrete elements of the simulated data $\boldsymbol{d}_{it}$, which then translate into discrete changes in $\breve{\boldsymbol{\varphi}}(\boldsymbol{\theta})$ and therefore $C(\boldsymbol{\theta})$. As a result, gradient-based methods cannot be used to minimize $C(\boldsymbol{\theta})$, so researchers need to resort to much slower derivative-free optimization methods, which limits the complexity of models that can be estimated. Furthermore, non-smoothness of $C(\theta)$ also makes it more difficult to establish the convergence of any minimization algorithm, and to produce an estimate of the (asymptotic) variance of $\hat{\boldsymbol{\theta}}^{II}$.

2.2 Generalized Indirect Inference (GII)

The Generalized Indirect Inference (GII) method of Bruins et al. (2018) addresses this problem by replacing $\breve{\boldsymbol{\varphi}}(\boldsymbol{\theta}) = \frac{1}{M}\sum_{m=1}^{M} \breve{\boldsymbol{\varphi}}^m(\boldsymbol{\theta})$ with the smooth analogue $\tilde{\boldsymbol{\varphi}}(\boldsymbol{\theta}, \rho) = \frac{1}{M}\sum_{m=1}^{M} \tilde{\boldsymbol{\varphi}}^m(\boldsymbol{\theta}, \rho)$, where ρ is a smoothing parameter such that if the sample size grows large and $\rho \to 0$, $\tilde{\boldsymbol{\varphi}}(\boldsymbol{\theta}, \rho)$ converges in probability to the non-stochastic binding function $\boldsymbol{\varphi}(\boldsymbol{\theta})$. Furthermore

$$\hat{\boldsymbol{\theta}}^{GII} = \arg\min_{\boldsymbol{\theta}} (\tilde{\boldsymbol{\varphi}}(\boldsymbol{\theta}, \rho) - \hat{\boldsymbol{\varphi}})' \mathcal{W} (\tilde{\boldsymbol{\varphi}}(\boldsymbol{\theta}, \rho) - \hat{\boldsymbol{\varphi}}). \tag{4}$$

[7] Other papers using this weight matrix include Blundell et al. (2016), Morten (2019) and Blundell et al. (2021).

will converge in probability to the vector of true structural parameters $\boldsymbol{\theta}_0$ as the sample size grows large and $\rho \to 0$.

A straightforward method of constructing $\tilde{\boldsymbol{\varphi}}^m(\boldsymbol{\theta}, \rho)$ is to replaces all discrete variables in both the structural model and in the auxiliary model estimated on the simulated data by smooth analogues. In the generic model outlined above, instead of using (1) to determine simulated choices, they would be determined by

$$\tilde{d}^m_{itk}(\rho) = \mathcal{K}_\rho\left(\boldsymbol{v}^m_{itk}\right) \tag{5}$$

where each element of $\boldsymbol{v}^m_{itk}$ is given by $u^m_{itk} - u^m_{itj}$, and j ranges over all $j \in \{1, \ldots, K\}$ such that $j \neq k$; $\mathcal{K}_\rho(\boldsymbol{v})$ is a smoothing kernel such that $\mathcal{K}_\rho(\boldsymbol{v}) = \mathcal{K}(\boldsymbol{v}/\rho)$ and $\mathcal{K} : \mathbb{R}^{K-1} \to \mathbb{R}$ is a smooth multivariate cdf.

The standard choice for $\mathcal{K}$ is the multivariate logistic cdf $\mathcal{K}(v) = 1/(1 + \sum_{j=1}^{K-1} \exp(v_j))$. This cdf is twice continuously differentiable and has the advantage that the resulting smoothed choices obey the 'adding-up property' $\sum_{k=1}^{K} \tilde{d}^m_{itk}(\rho) = 1$ regardless of the value of ρ. While not required for the consistency of the procedure, the latter property is useful in practice, as it reduces the bias of common simulated moments such as choice shares when $\rho > 0$.

Two difficulties arise for this conventional approach to GII estimation in models where discrete outcomes affect the state variables of the model, as is common in dynamic discrete choice models. The first difficulty, noted in Altonji, Smith Jr and Vidangos (2013) and Bruins et al. (2018), is that large smoothing-induced biases can arise as each period's choice will inherit the smoothing bias from past choices, so biases accumulate as an individual moves through the model. Second, in models that feature forward-looking agents, solving the model becomes prohibitively difficult, as smoothing leads to a continuous state space.

One practicable way of dealing with this problem is to rely on a 'one-step-ahead' simulation instead of a full 'path' simulation for each simulated individual. With this approach, the state vector for each simulated individual is updated according to the choices of actual individuals in the data instead of their (smoothed) simulated choices. Then the state space remains discrete and no smoothing bias accumulates across periods.

However, a severe limitation of this approach is that it cannot be used in models with unobserved heterogeneity. One illustrative example is unobserved type heterogeneity, where the unobserved types are known to the agents but not to the econometrician. The problem in

that case is that it is generally impossible to reliably assign the unobserved types to individuals in the data. Assigning unobserved types at random according to their *ex ante* probability will lead to bias. The reason is that agents choose their paths through the model *based on* knowledge of their unobservable type, so individuals with certain choice histories will be selected by type.[8]

As an example, consider a dynamic labour force participation model in the spirit of Eckstein and Wolpin (1989), where agents choose each period between market work and home production. Assume that there are two types of agents: some have a strong preference for market work and some for home production. Suppose one of the moments used to estimate the model is the share of agents transitioning to market work after a spell of home production. In the data, this share might be quite small, as agents in home production in any given period are likely to be disproportionately those with a strong preference for home production. In a one-step-ahead simulation at the true parameters, with types assigned according to their *ex ante* probability, the share transitioning to market work would likely be larger, as a proportional share of those in home production in any given period would be assigned strong preferences for market work.

A different approach, which we pursue in this paper, is to compute latent variables using all *possible* values of relevant past discrete outcomes, and then use a weighting procedure to generate a smoothed discrete outcome for the current period. A minimal example, discussed in Appendix A.1 of Bruins et al. (2018), is a discrete choice model with $K = 2$ where only the choice in $t-1$ matters for the choices in t. Then smooth choices in period t can be obtained recursively using

$$\widetilde{\widetilde{d}}^m_{itk}(\rho) = \widetilde{\widetilde{d}}^m_{i,t-1,1}(\rho)\mathcal{K}_\rho\left[\boldsymbol{v}^m_{itk}(d_{i,t-1,1}=1)\right] + \widetilde{\widetilde{d}}^m_{i,t-1,2}(\rho)\mathcal{K}_\rho\left[\boldsymbol{v}^m_{itk}(d_{i,t-1,2}=1)\right]. \tag{6}$$

where $d_{i,t-1,k} = 1$ is shorthand for a choice $\boldsymbol{d}_{i,t-1}$ in the previous period such that the kth element was chosen.

2.3 GII for complex dynamic discrete choice models

We adapt this approach for models where *all* past choices matter for current period choices, as is typically the case in the finite-horizon dynamic discrete choice models that are common in

[8] The same issue also arises for other kinds of unobserved states. An alternative approach of performing separate simulations for each type and then weighting by the *ex ante* probability of each type is subject to the same problem.

labour economics. Smooth choices are obtained using

$$\widetilde{\widetilde{d}}_{itk}^{m}(\rho) = \sum_{\mathfrak{d}_{i,t-1} \in D_{t-1}} W_{it}^{m}(\mathfrak{d}_{i,t-1}, \rho) \mathcal{K}_{\rho}\left[v_{itk}^{m}(\mathfrak{d}_{i,t-1})\right] \tag{7}$$

where D_{t-1} is the set of all possible sequences of discrete choices agents could have taken through the model up to the beginning of period t.[9] $W(\mathfrak{d}_{i,t-1}, \rho)$ is the weight term

$$W_{it}^{m}(\mathfrak{d}_{i,t-1}, \rho) = \begin{cases} W_{i,t-1}^{m}(\mathfrak{d}_{i,t-2}, \rho) \frac{\sum_{k=1}^{K} d_{i,t-1,k} \mathcal{K}_{\rho}\left[v_{i,t-1,k}^{m}(\mathfrak{d}_{i,t-2})\right]}{\sum_{k=1}^{K} \mathcal{K}_{\rho}\left[v_{i,t-1,k}^{m}(\mathfrak{d}_{i,t-2})\right]} & \text{if } t > 1 \\ 1 & \text{if } t = 1. \end{cases} \tag{8}$$

where $d_{i,t-1,k}$ is the kth element of $\boldsymbol{d}_{i,t-1}$, which in turn is the last element of $\mathfrak{d}_{i,t-1}$.

The framework given by (7) and (8) extends the technique proposed in Bruins et al. (2018) to a setting where there is an arbitrary number of options K and the whole path of previous choices $\mathfrak{d}_{i,t-1}$ matter for current choices. Division by $\sum_{k=1}^{K} \mathcal{K}_{\rho}\left[v_{i,t-1,k}^{m}(\mathfrak{d}_{i,t-2})\right]$ in (8) is optional but will improve the small sample properties of the estimator for smoothing kernels that do not guarantee $\sum_{k=1}^{K} \mathcal{K}_{\rho}\left[v_{i,t-1,k}^{m}(\mathfrak{d}_{i,t-2})\right] = 1$. (6) is nested as a special case provided that this latter equality holds, which is guaranteed for the standard choice for $\mathcal{K}$, the multivariate logistic cdf.[10]

While it would be possible to stop here and simply replace all $\boldsymbol{d}_{it}$ by $\widetilde{\widetilde{\boldsymbol{d}}}_{it}^{m}(\rho)$ according to (7) in the auxiliary model, it is usually preferable to extend this reweighting approach to the auxiliary model itself. For instance, continuous outcome variables that depend on past discrete outcomes can be calculated by averaging over their values given all *possible* values of relevant past discrete outcomes as in (7), with the weights given by (8). Similarly, a linear regression model relating current to past outcomes might be smoothed by weighting according to (8). How exactly this is done is best demonstrated by example; we show how to implement the approach for the model of Keane and Wolpin (1994) in section 3.2 and for the main application discussed in this paper in section 4.4 and Appendix C.

An intuitive interpretation of our procedure relies on the notion of a 'close counterfactual'.

[9] Throughout this section, $v_{itk}^{m}(\mathfrak{d}_{i,t-1})$ refers to v_{itk}^{m} calculated given the choice history $\mathfrak{d}_{i,t-1}$.

[10] More precisely, (7) and (8) nest

$$\widetilde{\widetilde{d}}_{itk}^{m}(\rho) = \begin{cases} \widetilde{\widetilde{d}}_{i,t-1,1}^{m}(\rho) \mathcal{K}_{\rho}\left[v_{itk}^{m}(d_{i,t-1,1} = 1)\right] + \widetilde{\widetilde{d}}_{i,t-1,2}^{m}(\rho) \mathcal{K}_{\rho}\left[v_{itk}^{m}(d_{i,t-1,2} = 1)\right] & \text{if } t > 1 \\ \mathcal{K}_{\rho}\left(v_{itk}^{m}\right) & \text{if } t = 1, \end{cases}$$

which is a variant of (6) with a defined first period at $t = 1$, provided that $\sum_{k=1}^{K} \mathcal{K}_{\rho}\left[v_{i,t-1,k}^{m}(\mathfrak{d}_{i,t-2})\right] = 1$.

For each individual in the structural model, our smoothing procedure assigns positive weight not only to her actual path through the model, but also to paths that are 'close counterfactuals' in the sense of having almost been taken. In other words, our simulation sample includes each simulated individual's *doppelgangers* in 'close possible worlds' as well, weighted according to their closeness. All paths are included in the auxiliary model, weighted by a measure of their proximity to the actual path. As the parameter vector changes in estimation, different paths are smoothly reweighted.

In order to make this approach computationally feasible, it is necessary to use a smoothing kernel with compact support in (8), so that non-zero weight is assigned to only some of the possible combinations of the values of the discrete variables. The reason is that the set D_{t-1} of possible paths through the model up to time t grows exponentially with t. As a result, unless most paths can be disregarded, the summation in (7) will usually become impossible to compute for reasonably large values of t. For instance, the canonical model of Keane and Wolpin (1994) has $K = 4$ choices and $T = 40$ periods, so the set D_{T-1} has $K^{T-1} \approx 3 \times 10^{23}$ elements.

When discrete outcomes are binary, a convenient choice of smoothing kernel with compact support is

$$\mathcal{K}_1(v) = \begin{cases} 0 & \text{for } v \leq -1 \\ \frac{1+v(2+v)}{2} & \text{for } -1 < v \leq 0 \\ \frac{1+v(2-v)}{2} & \text{for } 0 < v < 1 \\ 1 & \text{for } v \geq 1. \end{cases} \tag{9}$$

For outcomes with three or more options, we use

$$\mathcal{K}_2(v_1, \ldots, v_K) = \begin{cases} \frac{1}{1+\sum_{j=1}^{K-1} \exp(v_j)} & \text{for } 1 + \sum_{j=1}^{K-1} \exp(v_j) < \phi \\ \frac{\left[\left[1+\sum_{i=1}^{K-1} \exp(v_j)\right]-3\phi\right]^2}{4\phi^3} & \text{for } \phi \leq 1 + \sum_{j=1}^{K-1} \exp(v_j) < 3\phi \\ 0 & \text{for } 3\phi < 1 + \sum_{j=1}^{K-1} \exp(v_j). \end{cases} \tag{10}$$

$\mathcal{K}_1 : \mathbb{R} \to \mathbb{R}$ is closely related to the polynomial kernel recommended in Hajivassiliou, McFadden and Ruud (1996). $\mathcal{K}_2 : \mathbb{R}^{K-1} \to \mathbb{R}$ is a modification of the familiar Logit kernel, which switches to a polynomial form for values of the Logit denominator $1 + \sum_{j=1}^{K-1} \exp(v_j)$ that are

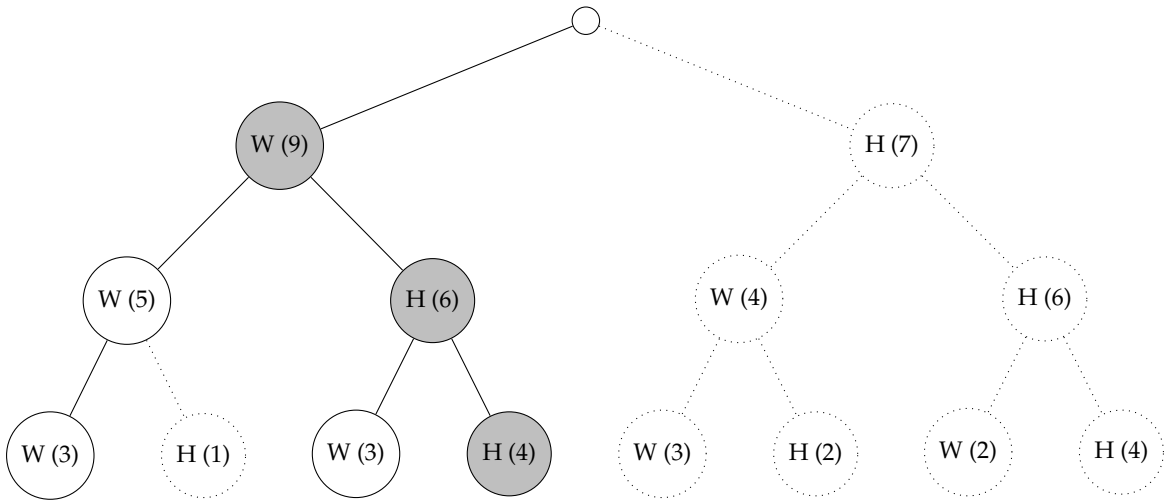
larger than the parameter $\phi > K/3$. Both kernels are continuously differentiable.[11]

It is important to note that it is possible to use a non-compact kernel in (7) and a compact kernel in (8). In section 3, we use the Logit kernel in (7) and $\mathcal{K}_2$ in (8) only. This has the advantage that current period choices retain the 'adding up property' that $\sum_{k=1}^{K} \mathcal{K}_\rho \left[\boldsymbol{v}_{i,t-1,k}(\mathfrak{d}_{i,t-1}) \right] = 1$ regardless of the value of ρ.[12]

2.4 Illustration: dynamic choice with $K = 2$ options and $T = 3$ periods

This subsection illustrates the different versions of the GII method using a simple example. Consider again a dynamic labour force participation model in the spirit of Eckstein and Wolpin (1989), this time without unobserved heterogeneity. Again, suppose agents choose each period between market work (W) and home production (H). Assume the model can be written as a particular specification of the generic model outlined above with $K = 2$ options and $T = 3$ periods.

Figure 1: Illustrative path through the model



Notes. This graph illustrates the possible path through this model for a particular individual. An explanation is given in the text.

Figure 1 illustrates the possible paths through this model for a particular individual. There are $K^T = 8$ possible paths the agent can take through the model. The associated utility u_{itk} for

[11] We prefer $\mathcal{K}_2$ to a purely polynomial kernel like $\mathcal{K}_3(v_1, \ldots, v_K) = \prod_{j=1}^{K-1} K_1(-v_j)$ because it exhibits an 'adding-up' property for large ϕ (inherited from Logit): $\sum_{k=1}^{K} K_2\left(\boldsymbol{v}_{itk}\right) \to 1$ as $\phi \to \infty$. A secondary advantage of this kernel is that the tuning parameter ϕ gives the econometrician a direct instrument for trading off the advantages of including relatively distant counterfactuals against the computational costs. $\phi > K/3$ is crucial to ensure $\sum_{k=1}^{K} \mathcal{K}_\rho \left[\boldsymbol{v}^m_{i,t-1,k}(\mathfrak{d}_{i,t-2}) \right] > 0$. The optimal choice of ϕ is likely to depend on ρ: the larger ρ, the smaller a ϕ will be required to retain computational feasibility.

[12] $\mathcal{K}_2$ only exhibits this adding up property as $\phi \to \infty$.

each option is given in parenthesis. The circles filled in light grey show the 'actual' path this simulated agent takes through the model: work in the first period and home production in the second and third period (WHH). Suppose utilities in this model depend on total past (market) work experience.

Some paths through the model are closer to this actual path than others, in the sense that (expected) utilities would need to change by less for the agent to have chosen them instead. With utilities as displayed in the diagram, there are two other paths that could be reached if utilities were changed by less than one unit each: WHW and WWW. These 'close counterfactual' paths are marked with solid lines in Figure 1; more distant counterfactual paths are marked with dotted lines.

Traditional GII in this model would obtain smooth simulated choices in each period according to (5), and use the same smoothed choices to update the individual's past work experience. This would have two undesirable consequences: expected utilities would need to be calculated for fractional years of work experience, and smoothing biases might accumulate across periods. This can be avoided using our procedure, which for $t = 2$ in this model can be written as

$$\widetilde{\widetilde{d}}^m_{i2k}(\rho) = \sum_{k_1=1}^{2} K_\rho \left[\boldsymbol{v}^m_{i2k}(d_{i1k_1} = 1) \right] \mathcal{K}_\rho(\boldsymbol{v}^m_{i1k_1}). \tag{11}$$

Again $d_{itk} = 1$ is shorthand for a choice $\boldsymbol{d}_{it}$ in period t such that the kth element was chosen.[13] A generalization to period $t = 3$ is

$$\widetilde{\widetilde{d}}^m_{i3k}(\rho) = \sum_{k_2=1}^{2} \sum_{k_1=1}^{2} \mathcal{K}_\rho \left[\boldsymbol{v}^m_{i3k}(d_{i1k_1} = 1, d_{i2k_2} = 1) \right] K_\rho \left[\boldsymbol{v}^m_{i2k_2}(d_{i1k_1} = 1) \right] K_\rho(\boldsymbol{v}^m_{i1k_1}). \tag{12}$$

While summing over all possible choice histories is practicable in this model, where the maximum number is $K^{T-1} = 4$, this is not practicable in more complex models with many choices and periods. We solve this problem by using a kernel with compact support like (9) or (10) and exploiting the recursive structure of the problem made explicit in (8). Using a compact kernel means that the weight of distant counterfactuals will be zero. For instance, suppose we used $\mathcal{K}_1$ given in (9) and set $\rho = 2$. With the utility values as given in Figure 1, that means only the 'actual' path through the model WHH and the close counterfactual paths WWW and WHH will have positive summands in (12).

[13] Division by $\sum_{k=1}^{K} \mathcal{K}_\rho \left[\boldsymbol{v}^m_{i,t-1,k}(\mathfrak{d}_{i,t-2}) \right]$ is omitted for ease of exposition.

Importantly, it is not necessary to compute all of the summands to find this out (which would be infeasible in a model with a large number of periods), as one choice not being 'close' rules out all choice histories containing it. In the given example, the option of home production in $t = 1$ is already not 'close' for the agent, so $\mathcal{K}_\rho(v^m_{i2k_1}) = 0$. This means that the whole branch of the decision tree starting with home production can be ignored in future calculations. In this case, it means that only two possible choice histories (WW and WH) need to be considered in period $t = 3$. in general, the recursive calculation of weights means that only close counterfactual paths need to be followed.

3 A test case: the model of Keane and Wolpin (1994)

3.1 The model

The model of Keane and Wolpin (1994) is a model of occupational choice. In each of $T = 40$ discrete periods of time, agents choose between $K = 4$ different options: blue-collar work, white-collar work, education, and home production. The per-period utility functions are

$$u_{1t} = w_{1t} = \exp(\alpha_{10} + \alpha_{11}s_t + \alpha_{12}x_{1t} - \alpha_{13}x_{1t}^2 + \alpha_{14}x_{2t} - \alpha_{15}x_{2t}^2 + \epsilon_{1t}) \tag{13}$$

$$u_{2t} = w_{2t} = \exp(\alpha_{20} + \alpha_{21}s_t + \alpha_{22}x_{2t} - \alpha_{23}x_{2t}^2 + \alpha_{24}x_{1t} - \alpha_{25}x_{1t}^2 + \epsilon_{2t}) \tag{14}$$

$$u_{3t} = \beta_0 - \beta_1 I(s_t \geq 13) - \beta_2(1 - d_{3,t-1}) + \epsilon_{3t} \tag{15}$$

$$u_{4t} = \gamma_0 + \epsilon_{4t} \tag{16}$$

w_{1t} and w_{2t} are the agent's (latent) wages in occupation one and two. s_t is the number of periods of schooling accumulated by the beginning of period t. x_{1t} and x_{2t} are an individual's total periods of work experience at the beginning of period t in occupation one and two, respectively. $\boldsymbol{d}_t$ is a vector of indicator variables, where for each element $d_{kt} = 1$ if option k is chosen in period t and $d_{kt} = 0$ otherwise. Hence $d_{3,t-1}$ is an indicator variable of whether schooling was chosen in the previous period. $\boldsymbol{\epsilon}_t \sim N(0, \Sigma)$ is a vector of serially uncorrelated shocks, where Σ is parameterized by the parameter vector $\boldsymbol{a}$. $\boldsymbol{\epsilon}_t$ is known to the agent at the beginning of period t but not before. $\boldsymbol{\theta} = \{\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \boldsymbol{\beta}, \gamma, \boldsymbol{a}\}$ is the full parameter vector of the model.

Individuals have perfect knowledge of the true model and rational expectations about the

future. Let $\boldsymbol{S}_t = \{s_t, x_{1t}, x_{2t}, d_{3,t-1}, \boldsymbol{\epsilon}_t\}$ be the vector of state variables, and denote the vector of pre-determined state variable by $\overline{\boldsymbol{S}}_t = \{s_t, x_{1t}, x_{2t}, d_{3,t-1}\}$. The vector of pre-determined state variables evolves deterministically given the previous period's choices: $x_{1,t+1} = x_{1,t} + d_{1,t}$, $x_{2,t+1} = x_{2,t} + d_{2,t}$, $s_{t+1} = s_t + d_{3,t}$, and $d_{3,t-1}$ is simply the previous period's choice. Initial conditions are $x_{11} = x_{21} = 0$ and $s_1 = 10$. All agents are *ex ante* homogeneous; different choices arise only due to different draws of $\boldsymbol{\epsilon}_t$.

The agent's decision problem in period t can be written recursively as

$$\underset{k \leq K}{\operatorname{argmax}} \, V_{kt}(\boldsymbol{S}_t) \tag{17}$$

where

$$V_{kt}(\boldsymbol{S}_t) = \begin{cases} u_{kt}(\boldsymbol{S}_t) + \delta \, \mathrm{E}\left[\max_{j \leq K} V_{j,t+1}(\boldsymbol{S}_{t+1}) | \overline{\boldsymbol{S}}_t, d_{kt} = 1\right] & \text{if } t < T \\ u_{kt}(\boldsymbol{S}_t) & \text{if } t = T \end{cases} \tag{18}$$

is the *alternative-specific value function*. E denotes the mathematical expectations operator. The discount factor δ is fixed at $\delta = 0.95$.

3.2 Estimating the Keane and Wolpin (1994) model

The agent's problem in this model can be solved by backward recursion using (17) and (18). The expectation in (18) can be evaluated using Monte Carlo integration.[14] Various methods can be employed to estimate the model, as the likelihood function is tractable and there is no unobserved heterogeneity. Here we compare three methods: Simulated Maximum Likelihood as used in Keane and Wolpin (1994), conventional GII with a one-step-ahead forecast, and the extended GII method outlined in section 2.3.

Simulated Maximum Likelihood. Our Simulated Maximum Likelihood procedure closely follows Keane and Wolpin (1994). The log likelihood function takes the form

$$l(\theta) = \sum_{i=1}^{N} \left[\log f(\boldsymbol{d}_{i1}, w_{i1}) + \sum_{t=2}^{T} \log f(\boldsymbol{d}_{it}, w_{it} | (\boldsymbol{d}_{i,s})_{s=1}^{s=t-1}) \right]. \tag{19}$$

where $f(\boldsymbol{d}_{it}, w_{it})$ is the joint density of agent i's choice in period t and the corresponding observed wage. If no wage is observed, because the agent chose schooling ($k = 3$) or home pro-

[14] Alternative solution methods are discussed in chapter 3 of this thesis.

duction ($k = 4$), $f(\boldsymbol{d}_{it}, w_{it})$ is simply the probability of the observed choice $P(\boldsymbol{d}_{it})$ conditional on the observed state variables.

We simulate the density $f(\boldsymbol{d}_{it}, w_{it})$ using the simulation technique proposed in chapter 1 of this thesis, which in the absence of measurement error in wages is numerically equivalent to the procedure likely used in Keane and Wolpin (1994).[15] The probabilities that enter the density are smoothed using the Logit smoother first proposed by McFadden (1989).[16] This is necessary to avoid numerical problems related to simulated probabilities of zero; it has the important added advantage that it leads to a smooth log-likelihood function and thus permits the use of gradient-based optimization routines.[17]

GII with a one-step-ahead forecast. This model is somewhat more complex than the prototypical model outlined in section 2, as in addition to a discrete choice d_{it}, a continuous wage w_{it} is observed in periods where one of the market work options is chosen. Our approach falls into the subcategory of Simulated Method of Moments (SMM), as the parameters of our auxiliary model are moments of the data. These are: the share of individuals choosing each option in each period, the mean log wage in each market occupation in each period, the variance of log wages in each occupation in each period, and transition matrices from each choice to each other choice in each period.

As the model features no unobserved heterogeneity, it is possible to use 'one-step-ahead' simulations instead of full path simulations: the state vector is updated using the choices observed in the data rather than the simulated choices. Smooth simulated moments are then straightforward to obtain by replacing simulated choices $\boldsymbol{d}_{it}$ with smooth choices $\tilde{\boldsymbol{d}}_{it}^m$ obtained using (5) in the calculation of moments. In particular, smooth choice shares are

$$\widetilde{c_{tk}^m} = \frac{1}{N}\sum_{i=1}^{N} \mathcal{K}_\rho(\boldsymbol{v}_{itk}) = \frac{1}{N}\sum_{i=1}^{N} \tilde{d}_{itk}^m(\rho). \tag{20}$$

Smooth means and variance can be calculated as

$$\widetilde{\log w_{tk}^m} = \frac{\sum_{i=1}^{N} \tilde{d}_{itk}^m(\rho) \log w_{itk}^m}{\sum_{i=1}^{N} \tilde{d}_{itk}^m(\rho)} \tag{21}$$

[15] For details, see chapter 1.

[16] Like Keane and Wolpin (1994), we use a tuning parameter of 500.

[17] We use the BHHH algorithm (Berndt et al., 1974), as extensive experimentation showed this algorithm to be much more efficient than other commonly used algorithms such as Simplex, Simulated Annealing, or BFGS.

and

$$\widetilde{\text{Var}}(\log w_{tk}^m) = \frac{\sum_{i=1}^N \tilde{d}_{itk}^m(\rho) \left[\log w_{itk}^m\right]^2 - \sum_{i=1}^N \tilde{d}_{itk}^m(\rho) \left[\widetilde{\log w_{tk}^m}\right]^2}{\left[\sum_{i=1}^N \tilde{d}_{itk}^m(\rho)\right] - 1}. \tag{22}$$

The (j,k) entry in the smoothed transition matrix from $t-1$ to t can be obtained as

$$\widetilde{M_{tjk}^m} = \frac{\sum_{i=1}^N \tilde{d}_{itk}^m(\rho) d_{i,t-1,j}}{\sum_{i=1}^N d_{i,t-1,j}}. \tag{23}$$

Extended GII. With this method, the state vector is updated according to simulated choices rather than observed choices, which makes the method suitable for models with unobserved heterogeneity. Then it is necessary to keep track not only of simulated agents' actual paths through the model, but also of paths that are 'close counterfactuals'. Smooth choice shares can be constructed using (7):

$$\widetilde{\widetilde{c_{tk}^m}} = \frac{1}{N} \sum_{i=1}^N \sum_{\mathfrak{d}_{i,t-1} \in D_{t-1}} K_\rho \left[v_{itk}^m(\mathfrak{d}_{i,t-1})\right] W_{it}^m(\mathfrak{d}_{it}, \rho) = \frac{1}{N} \sum_{i=1}^N \widetilde{\widetilde{d_{itk}^m}}(\rho). \tag{24}$$

In order to obtain smoothed means and variances of log wages, one option would be to use past values of $\widetilde{\widetilde{d_{it}^m}}(\rho)$ to construct state variables. While the resulting wage moments would be smooth and converge to the unsmoothed wage moments as $\rho \to 0$, they would inherit the smoothing biases of choices in all past periods, and may therefore exhibit large biases for $\rho > 0$. We therefore prefer to construct the wage moments by analogy to (7), in keeping with the heuristic of treating all paths with positive weight as 'possible' paths through the model.

$$\widetilde{\widetilde{\log w_{tk}^m}} = \frac{\sum_{i=1}^N \sum_{\mathfrak{d}_{i,t-1} \in D_{t-1}} K_\rho \left[v_{itk}^m(\mathfrak{d}_{i,t-1})\right] W_{it}^m(\mathfrak{d}_{i,t-1}, \rho) \log w_{itk}^m(\mathfrak{d}_{i,t-1})}{\sum_{i=1}^N \widetilde{\widetilde{d_{it}^m}}(\rho)} \tag{25}$$

where weights are constructed according to (8). Smooth variances of log wages are constructed analogously as

$$\widetilde{\widetilde{\text{Var}}}(\log w_{tk}^m) = \frac{\sum_{i=1}^N \sum_{\mathfrak{d}_{i,t-1} \in D_{t-1}} K_\rho \left[v_{itk}^m(\mathfrak{d}_{i,t-1})\right] W_{it}^m(\mathfrak{d}_{i,t-1}, \rho) \left[\log w_{itk}^m(\mathfrak{d}_{i,t-1})\right]^2 - \sum_{i=1}^N \widetilde{\widetilde{d_{itk}^m}}(\rho) \left[\widetilde{\widetilde{\log w_{tk}^m}}\right]^2}{\left[\sum_{i=1}^N \widetilde{\widetilde{d_{itk}^m}}(\rho)\right] - 1} \tag{26}$$

By the same logic, we construct the (j,k) entry in the smoothed transition matrix from $t-1$ to t as

$$\widetilde{\widetilde{M_{tjk}^m}} = \frac{\sum_{i=1}^N \sum_{\mathfrak{d}_{i,t-1} \in D_{t-1}} K_\rho \left[v_{itk}^m(\mathfrak{d}_{i,t-1})\right] W_{it}^m(\mathfrak{d}_{i,t-1}, \rho) d_{i,t-1,j}(\mathfrak{d}_{i,t-1})}{\sum_{i=1}^N \sum_{\mathfrak{d}_{i,t-1} \in D_{t-1}} W_{it}^m(\mathfrak{d}_{i,t-1}, \rho) d_{i,t-1,j}(\mathfrak{d}_{i,t-1})} \tag{27}$$

where $d_{i,t-1,j}(\mathfrak{d}_{i,t-1})$ refers to the choice $d_{i,t-1,j}$ as implied by the possible choice history $\mathfrak{d}_{i,t-1}$.

3.3 Monte Carlo results.

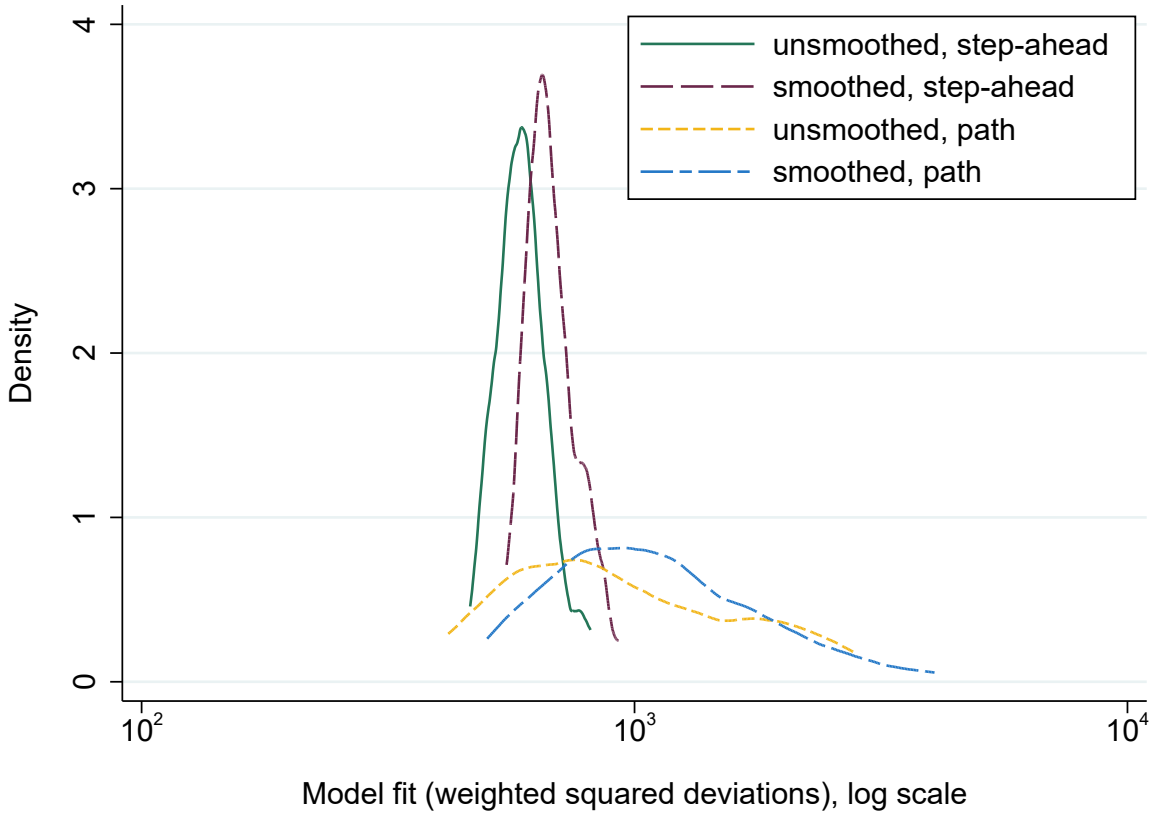
Approximation performance. We use simulations from the model evaluated at the true parameters $\boldsymbol{\theta}_0$ to assess how closely the smoothed simulated moments (or in the language of Indirect Inference, the parameters of the smoothed auxiliary model estimated from the simulated data) align with the data moments (or the parameters of the auxiliary model estimated from the actual data), compared to the unsmoothed simulated moments.[18] For this purpose, we simulated 100 'actual' data sets with $N = 10,000$ individuals observed for $T = 40$ periods each, with the model solution as close as possible to the truth (based on $D = 80,000$ simple random draws in Monte Carlo simulation).

We first record the distances between the (unsmoothed) moments obtained from these 'actual' data sets and the *unsmoothed* simulated moments obtained from simulating another 100 data sets from the model at the true parameters, with $D = 2000$ simulation draws and using a systematic sampling method. We perform this step both using path simulation and a one-step-ahead simulation, whereby state variables are updated according to choices observed in the data rather than simulated choices. We then obtain *smoothed* simulated moments using the extended GII method and a path simulation and the conventional GII method and a one-step-ahead simulation. The distances are measured using the criterion functions given by (2) and (4).

Figure 2 visualizes the results of this exercise. For each of the four simulation options (unsmoothed path, unsmoothed one-step-ahead, smoothed path, smoothed one-step-ahead), we show the kernel-estimated density of the value of the criterion function. Unsurprisingly, the one-step-ahead simulations yield a better fit, as differences in choice shares in one period do not feed through to later periods. For both one-step-ahead and path simulations, smoothed moments are only slightly further away from the data moments than unsmoothed moments.

[18] More specifically, we use the parameter values from 'data set one' in Keane and Wolpin (1994), listed in Table 1 of that paper (and in Appendix D of chapter 3 of this thesis).

Figure 2: Approximation Performance



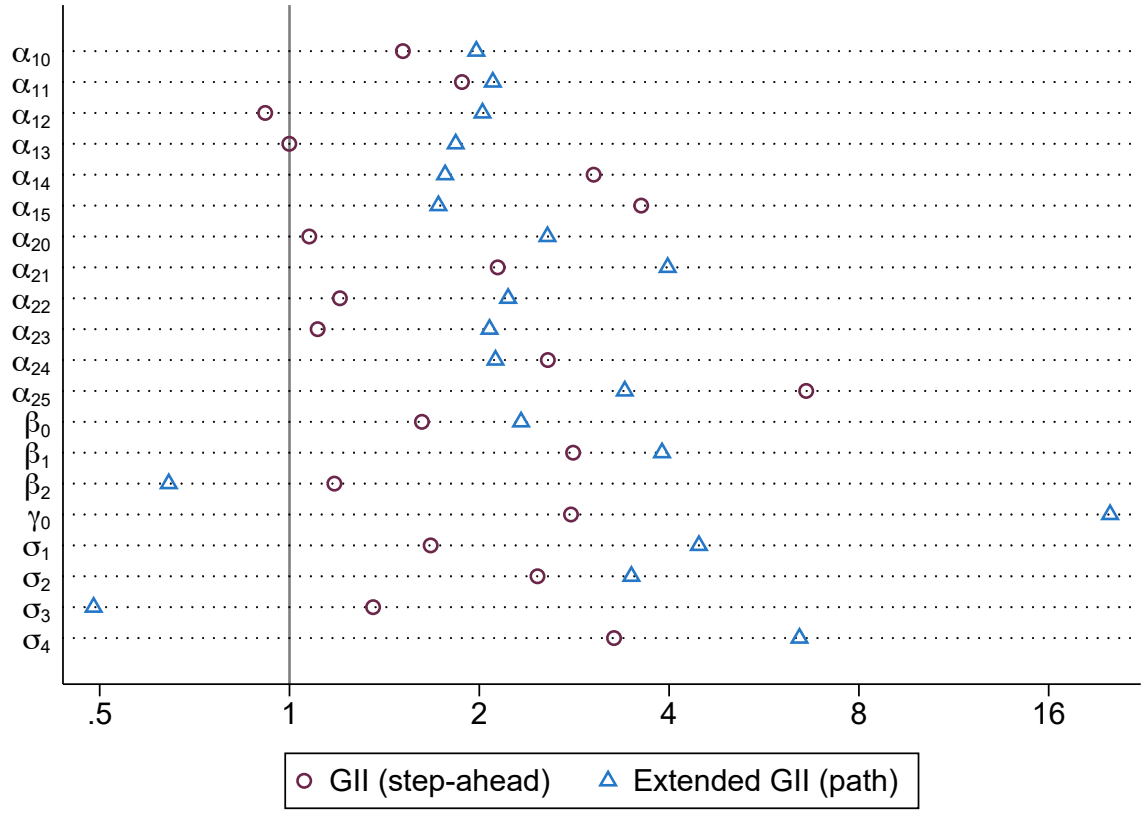
Notes. The densities shown are kernel densities estimated using an Epanechnikov kernel with the bandwidth selected using Silverman's Rule. In each case, the kernel density is calculated over 100 approximation runs with different shocks and approximation draws. All draws use systematic sampling.

Estimation Performance. We estimate the model using data from $N = 2000$ individuals for Simulated Maximum Likelihood and $N = 10,000$ individuals for the Simulated Method of Moments. When using maximum likelihood, 200 Halton draws were used to simulate the likelihood contribution of each observation. For the Simulated Method of Moments, we set the number of simulated samples to $M = 1$. Thirty estimation runs were performed; the six estimation runs (20 percent) with the lowest simulated likelihood at the estimated parameters were dropped to guard against outliers and numerical problems.

As in Keane and Wolpin (1994), all of the results are presented for a fixed number of numerical integration draws for both solution and estimation, and with smoothing parameters held constant. The motivation for this is that it makes our method more transparent and the results easier to interpret. In actual estimation, it would be advisable to increase the number of integration draws and reduce the amount of smoothing as the algorithm converges.[19]

[19]One way of implementing this would be as a two-step procedure, where (at least) one final step of a (Quasi-

Figure 3: Root mean squared error relative to full solution



Notes. The diagram shows the relative root mean squared errors of the estimates for all model parameters. Root mean squared errors obtained using Simulated Maximum Likelihood are normalized to one, as indicated by the grey vertical line. Root mean squared errors were calculated over 24 sets of estimated parameters. These were obtained by estimating the model parameters 30 times from different simulated data sets; the 6 estimated parameter sets (20%) with, respectively, the lowest simulated likelihood or the highest value of the criterion function were dropped to guard against the effects of outliers and numerical problems. $N = 2000$ individuals were included for Simulated Maximum Likelihood and $N = 10,000$ individuals for the Simulated Method of Moments. $D = 2000$ solution draws were used in model solution. When using maximum likelihood, 200 Halton draws were used to simulate the likelihood contribution of each observation. For the Simulated Method of Moments, the number of simulated samples was set to $M = 1$.

In order to avoid spuriously positive results, we start the estimation procedures not at the true values but at a set of values that an empirical researcher without knowledge of the true parameter values might reasonably have picked as starting values. In particular, for the wage parameters, we use the parameters of two separate bi variate sample selection models estimated from the data.[20] We further set the value of home production γ_0 to one standard

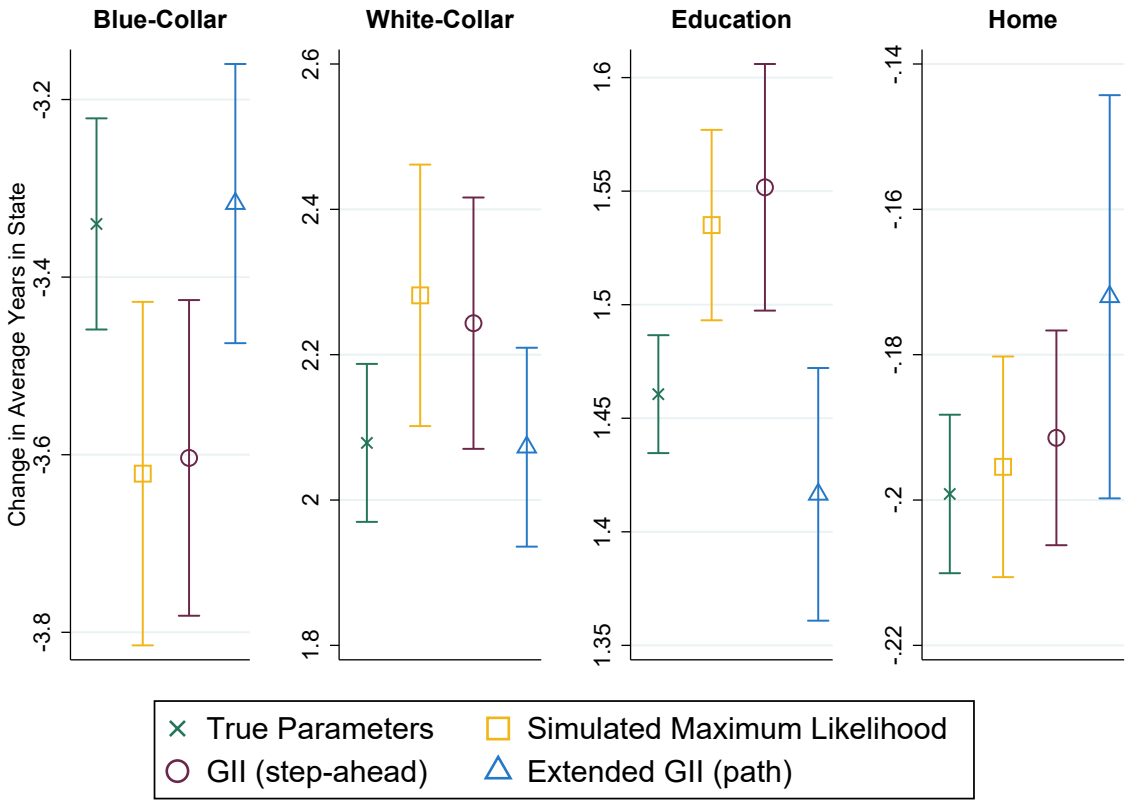
)Newton algorithm would be performed using the full solution, a higher number of draws, and less smoothing. This would likely yield additional accuracy gains at little computational cost. For theoretical arguments in favour of this two-step approach, see section 3.1 of Hajivassiliou (2000) and Kristensen and Salanié (2017). Kristensen and Salanié (2017) and Bruins et al. (2018) present Monte Carlo evidence suggesting that this approach also works well in practice.

[20]Two natural exclusion restrictions arise from the full model: it is known that which period an individual is in and whether school was chosen in the previous period can both affect continuation values and thus choices but not wages. We estimated the selection models using Maximum Likelihood (assuming normally distributed errors).

deviation below the median blue collar wage, and the standard deviation of the non-wage options to the average estimated standard deviation of the market-work options. All other parameters are set to zero.

Figure 3 shows the results of this exercise. Parameters estimates obtained using Simulated Maximum Likelihood (the computationally most expensive method) were generally closest to the true parameters. Both GII methods perform roughly equally well, with typical root mean squared errors around double those for Simulated Maximum Likelihood.

Figure 4: Estimated Impact of a $500 Tuition Subsidy



Notes. Estimated impact of a $500 tuition subsidy on average years spent in each occupation. Markers indicate the mean effect over 24 estimation runs; bars indicate one standard deviation. Estimation used balanced panels of $N = 10,000$ individuals for the Simulated Method of Moments and $N = 2000$ for Simulated Maximum Likelihood. Counterfactual simulations from the model are performed separately for each set of parameters by simulating paths for $10,000$ individuals.

Policy Experiment. Figure 4 shows how the performance of the three estimators translates into prediction accuracy in a simple policy experiment. The four panels show the estimated impact of a $500 tuition subsidy on the average number of years individuals spend in each

Using the Heckman (1979) two-step method would have resulted in starting values at a similar distance from the true parameters.

occupation in the Keane and Wolpin (1994) model. At the true parameters, the subsidy leads to a rise in the average number of years individuals spend in education and white-collar work, and a fall in the number of years spent in blue-collar work and home production.

Reassuringly, counterfactual simulations based on the estimated parameters are close to the results of simulations using the true parameters, regardless of which of the three estimation methods is used. Results using Generalized Indirect Inference and Simulated Maximum Likelihood are very close to one another; both slightly overestimate the magnitude of the effect. Sample standard deviations for all methods are similar.

Interestingly, predictions using the parameters estimated using the extended GII method are actually *closer* to prediction using the true parameters than those obtained using the other two methods (an exception is the effect on time spent in home production, but there the true effect is very small). This may be related to the beneficial side-effect of smoothing on sampling noise, as the extended GII method includes a number of 'close' paths through the model for each simulated individual. This likely increases the 'effective' sample size in simulation.

4 Empirical Application: Education Maintenance Allowance

In the OECD, one fifth of young adults do not complete upper secondary education by age 25 (OECD, 2017). Early school dropout is strongly correlated with poor subsequent labour market outcomes (Bäckman et al., 2015; Brunello and De Paola, 2014; Gregg and Tominey, 2005). A number of studies exploiting supply side variation suggest that the association is at least partly causal (see Card, 2001). Early dropout is therefore a major policy concern.

In this section, we use the method developed above to estimate a structural model of young people's career choices in the spirit of Eckstein and Wolpin (1999), with a particular focus on school dropout. Using this model, we investigate the labour market effects of the Education Maintenance Allowance (EMA), a conditional cash transfer (CCT) programme in the UK that paid young people up to £3,600 ($4,800) for staying in education or training between the ages of 16 and 19. A structural approach allows us to assess the long-run impact of the programme and enables us to identify the main mechanisms of action.

4.1 Institutional Background

Until 2014, school was compulsory in England up until GCSE examinations, which are taken at the end of the academic year in which individuals turn 16 (the academic year runs from September 1 until August 31). Since 2014, education or training has been technically compulsory between 16 and 18, although enforcement still typically only occurs until 16. In the mid 2000s, around 80% of students remained in education or training beyond 16 and around 70% beyond age 17, with some re-sitting GCSEs and others studying for university entrance qualifications.

The Education Maintenance Allowance (EMA) was a grant available to students from low-income households in 'non-advanced' (i.e. post-secondary but not tertiary) education between the ages of 16 and 19.[21] It was introduced nationwide in 2004 after a pilot scheme that began in 1999. The most generous form of the grant was worth £30 per week during term time, equivalent to £1,200 per year. Individuals with household incomes of less than £20,817 a year qualified for the full grant. Partial EMA grants worth £20 and £10 per week were available to individuals with household incomes below £25,521 and £30,810 per year, respectively. Notably, the grant was paid to students themselves, rather than to their parents.

The EMA was scrapped in England for the 2011/2012 academic year and replaced with the 'more targeted' 16-19 Bursary. Total spending on EMA grants was around £560 million, while the Bursary had a budget of £180 million. The Scottish Parliament, the Welsh Assembly and the Northern Irish Assembly rejected the policy, so the change occurred only in England. Unlike for the EMA, eligibility for the 16-19 Bursary is not clearly defined, with any individual who 'needs financial help' eligible to apply.[22] Individual schools then make the final decision on whether individuals get the grant.

In addition to the EMA, two other government transfers for post-secondary students were available in England during our sample period, both of which were paid to the parents. The first is Child Benefit, which in the relevant period was paid to the parents of any child age 19 or younger in 'non-advanced' education, and was worth around £18 per week for the eldest child and £12 per week for subsequent children. The second is the Child Tax Credit (CTC), which was also paid to parents of children in education age 19 or younger, but in contrast to Child Benefit was means-tested. The amount was dependent on parental income; parents

[21] More specifically, advanced education is education beyond A-Levels or equivalent.

[22] The 'Vulnerable Bursary' – set up as part of the 16-19 Bursary – does actually have strict eligibility criteria, but only affects a small number of individuals.

in our sample period received around £2400 per year if their combined annual income was less than around £15,000 per year and less if their income was higher. Most households were eligible for some amount of CTC; only households with annual incomes above around £58,000 were ineligible.

In this system, the subsidy for keeping a child in education was greatest for families on incomes of less than £15,000, who could gain around £90 per week in total government support, of which the EMA made up roughly one third. As a proportion of the total subsidy, the EMA was the most significant for households with annual incomes of around £20,000, for whom it made up more than half of total government support. The reason is that these families were eligible for the full rate of the EMA at £30, but only for a relatively small amount of CTC.

4.2 Data

We estimate our model using data from the *Next Steps* cohort study (University College London, 2021), formerly known as the Longitudinal Survey of Young People in England (LSYPE). *Next Steps* follows English teenagers born between September 1989 and August 1990 from the end of Year 9 (age 14) up to age 26. The survey includes rich background information on the participants and tracks their economic activities. A Special Licence version of the study (University College London, 2020) includes linked administrative schools data from the national Pupil Database (NPD); as detailed below, we use this link to extract data on student attainment on nationally standardized age 11 exams (Key Stage 2 SATs) and age 16 exams (GCSEs).

Next Steps currently has eight waves. The first seven were conducted every year from 2004 (age 14), while the eighth was conducted after a five year gap in 2016 (age 26). In the final wave, individuals were asked about their economic activity in each of the missing years, giving us a complete activity history. Unfortunately, there is no wage data for the missing years, so we do not observe earnings for individuals at ages 21 to 25. As we are interested in young people's choices for the period beyond the statutory school leaving age, we use data on young people from wave 4 onward, which was conducted at the end of the first post-compulsory year of schooling (age 17).

Table 1 shows the sample sizes by wave, conditional on key background characteristics such as parental income and gender being observed. Overall there are around 10,600 usable individual records in wave 4 of the survey, which represent the vast majority of the 11,801

individuals that participated in this wave.[23] The sample size gradually declines with each wave, with only around 60% of the sample remaining in wave 8. Attrition is somewhat higher among men than among women.

Table 1: Overall *Next Steps* Sample Sizes

	Age	Men	Women	Total
Wave 4	17	5,404	5,252	10,656
Wave 5	18	4,969	4,987	9,956
Wave 6	19	4,459	4,587	9,046
Wave 7	20	3,873	4,048	7,921
Retrospective	21-25	2,653	3,224	5,877
Wave 8	26	2,753	3,336	6,089

Notes. Unweighted sample sizes conditional on key background characteristics being observed.

A key variable in our model is parental income, from which we are able to determine eligibility for the EMA. We use parental income in wave 3 of the survey, which is the last year of compulsory school and the year in which EMA eligibility was usually assessed. Table 2 shows the number of individuals in each of the EMA entitlement groups by wave.[24]

Table 2: Sample sizes by EMA eligibility

	Age	£30/week	£20/week	£10/week	Ineligible	Total
Wave 4	17	4,159	1,104	1,125	4,268	10,656
Wave 5	18	3,777	1,031	1,056	4,092	9,956
Wave 6	19	3,358	933	968	3,787	9,046
Wave 7	20	2,810	815	878	3,418	7,921
Retrospective	21-25	2,025	616	643	2,593	5,877
Wave 8	26	2,131	643	660	2,655	6,089

Notes. Unweighted sample sizes conditional on key background characteristics being observed.

Table 2 shows that in wave 4, there are around 4,200 individuals in the fully eligible (£30/week) group and around 4,300 individuals in the ineligible group. There are around 1,100 individu-

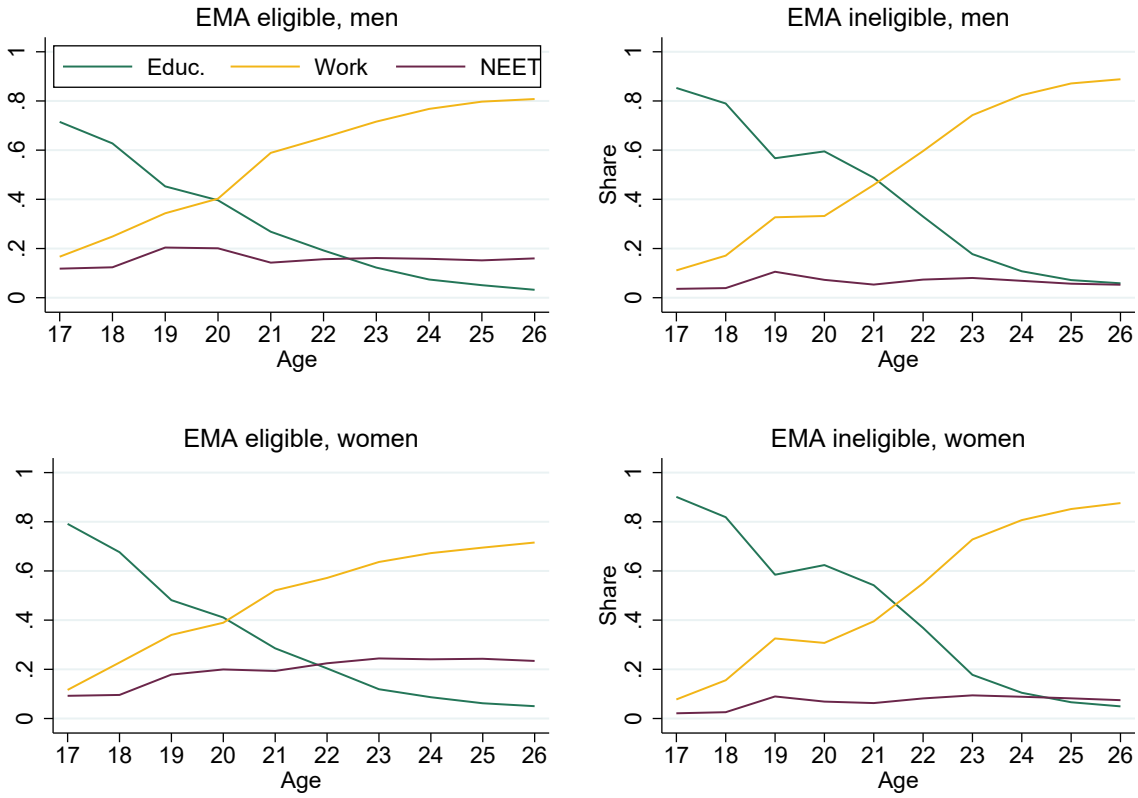
[23] Around a third of the non-usable records are accounted for by an 'ethnic minority boost sample' that was added in wave 4. We exclude these individuals, because we do not observe parental income for them in the relevant period.

[24] The parental income variable in the *Next Steps* study is banded. We use an interval regression of parental income on a rich set of covariates to generate an expected value for parent income conditional on falling within the reported band. Fortunately, the bands align closely with the EMA eligibility thresholds, so the number of wrongly assigned individuals is likely to be small.

als in each of the two partial eligibility (£20/week and £10/week) groups. The rate of attrition is slightly higher among the EMA eligible.

As the *Next Steps* study was conducted using a complex two-stage sampling design with disproportionate stratification (disadvantaged schools and ethnic minority pupils were initially oversampled), these unweighted sample sizes are not representative of the population. As a result, any analysis using *Next Steps* needs to use sampling weights to obtain a representative sample. Furthermore, all standard errors need to be adjusted to take into account the two-stage sampling design. Different weights are available for each wave to account for differential attrition. Using appropriate weights yields population shares of 36% in the fully eligible group, 10% in the £20/week group, 11% in the £10/week group, and 43% ineligible for the EMA.

Figure 5: Proportion in education, employment and NEET by age

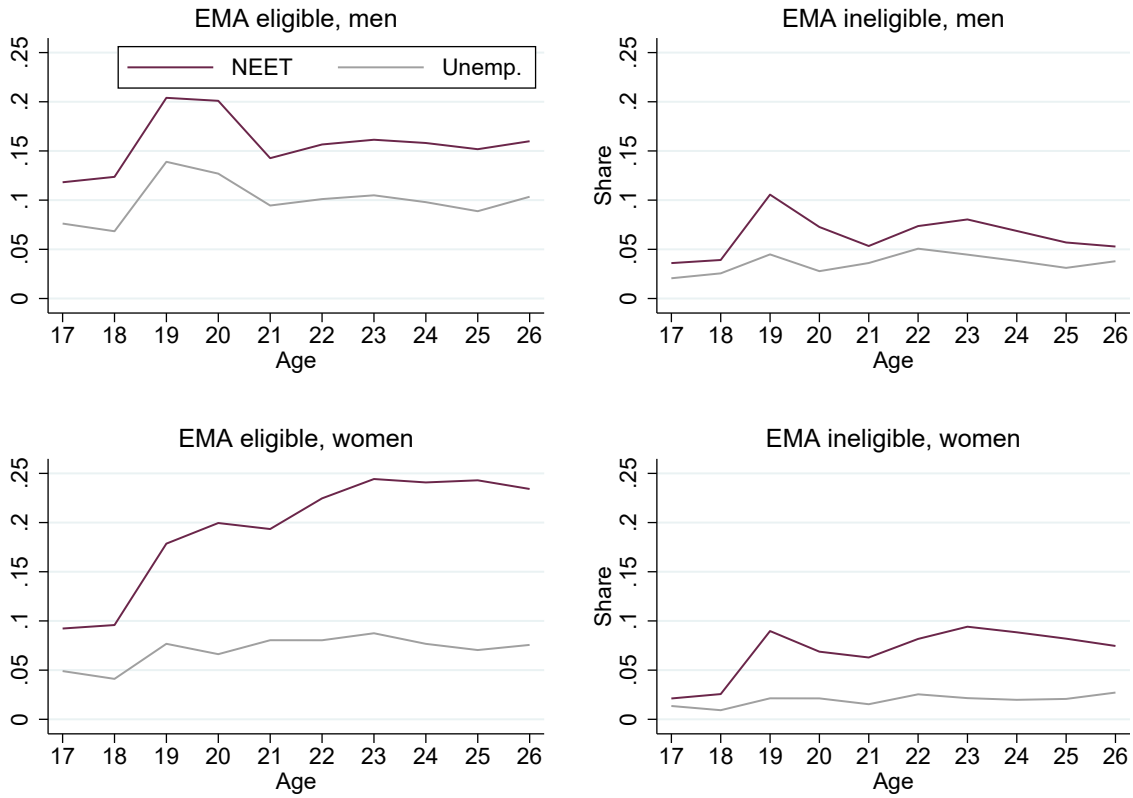


Notes. EMA eligibility is based on parental income at age 16 and includes the partially eligible (i.e. the £20 and £10 groups) as well as the fully eligible. 'Age' is the young person's age at the end of each period.

Figure 5 shows the (weighted) proportion of individuals in education or training, in work, and not in employment, education or training (NEET) by age. The sample is split by gender

and EMA eligibility. For all four groups there is a clear rise in the share in work and a decline in the share in education over time. There is a small *rise* in educational participation for EMA ineligible individuals between age 18/19 and age 19/20, which likely reflects the effect of gap years before university. Educational attendance in the early 20s and employment rates in the mid-20s are lower among those eligible for the EMA than among the ineligible. Differences by gender are mostly small, although the NEET share among EMA-eligible women is noticeably higher from age 21 onward.

Figure 6: Proportion of NEET and unemployed by age

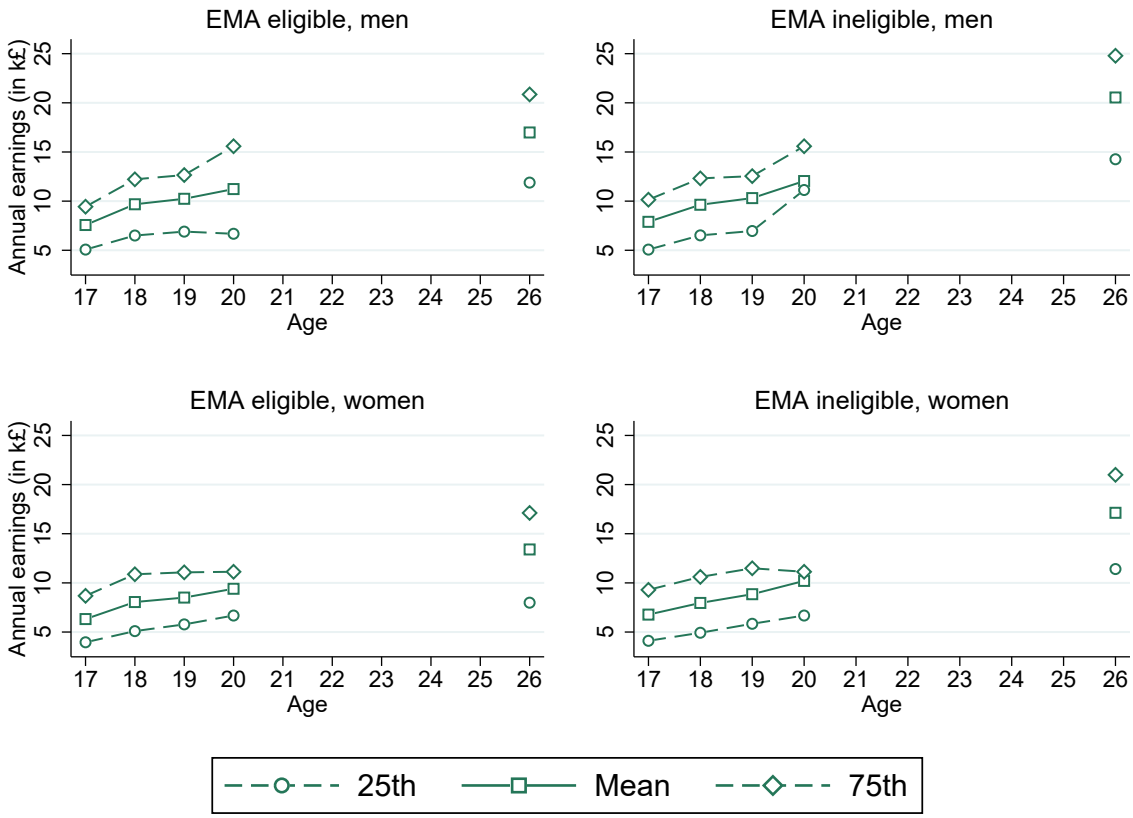


Notes. EMA eligibility is based on parental income at age 16 and includes the partially eligible (i.e. the £20 and £10 groups) as well as the fully eligible. 'Age' is the young person's age at the end of each period.

Figure 6 zooms in on the proportion of people not in education, employment or training (repeating the same curves from Figure 5) and adds the proportion who report being unemployed by age. The graphs show that across groups, around half of NEET individuals do not think of themselves as NEET by choice. For women, the share of NEET individuals who are unemployed is lower than it is for men, potentially reflecting the impact of motherhood. EMA eligible individuals are much more likely to be NEET and more likely to be unemployed at

every age.

Figure 7: Earnings distribution by Age

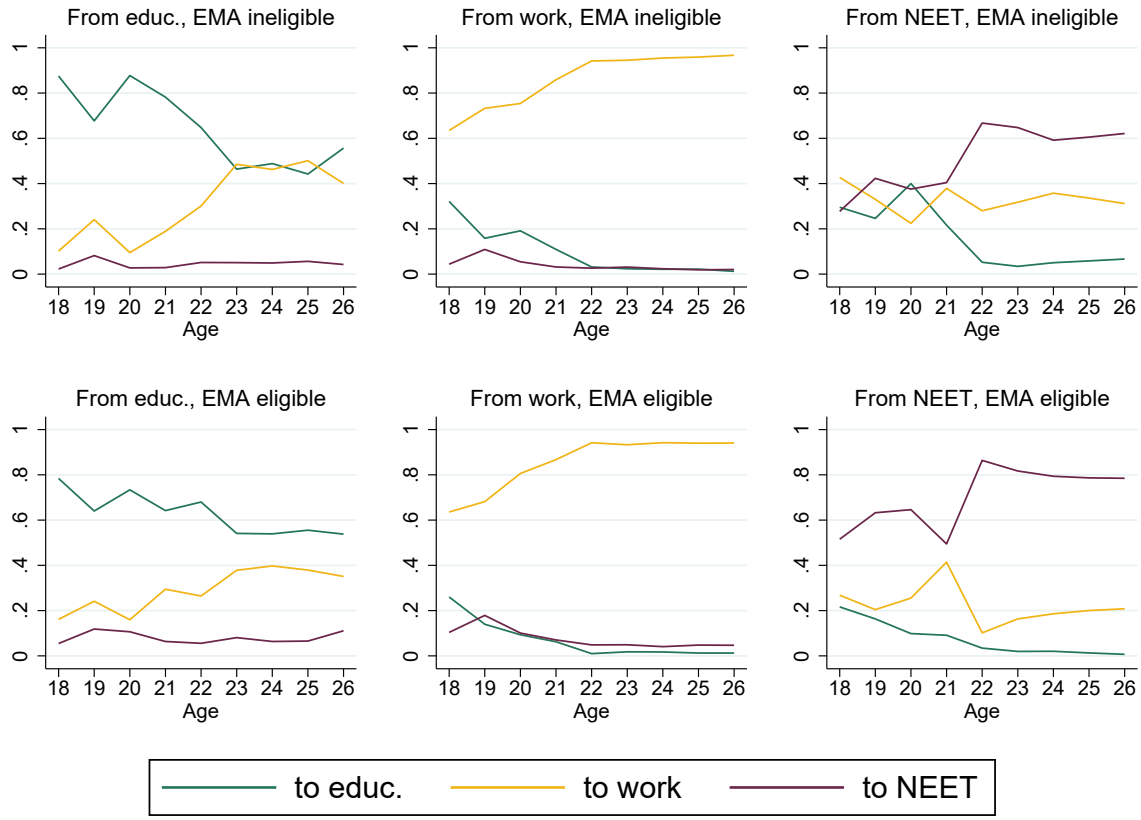


Notes. Mean, 25th and 75th percentiles of the earnings distribution for individuals whose main activity is work. Earnings are annualized and in 2006 prices. EMA eligibility is based on parental income at age 16. Wages are Winsorized at the 1st and 99th percentile. 'Age' is the young person's age at the end of each period.

Figure 7 shows the mean as well as the 25th and 75th percentiles of the earnings distribution by age for individuals whose main activity is work. To guard against outliers, earnings are Winsorized at the 1st and 99th percentile. As we do not observe earnings in between waves seven and eight of the *Next Steps* survey, there is no earnings data between 21 and 24 inclusive.[25] Earnings increase with age for all groups due to both earnings growth and the change in composition of the workforce: by age 25, the sample in work includes a much more educated set of individuals. Earnings are lower for those coming from poorer (EMA-eligible) households, and are lower for women than they are for men.

[25]Reported age 20 earnings are potentially less reliable than at other ages, as they are taken from a banded variable (the midpoints were assumed to reflect true earnings).

Figure 8: Transitions by Age



Notes. The diagram shows the share moving into each activity conditional on their activity in the previous year. EMA eligibility is based on parental income at age 16. 'Age' is the young person's age at the end of each period.

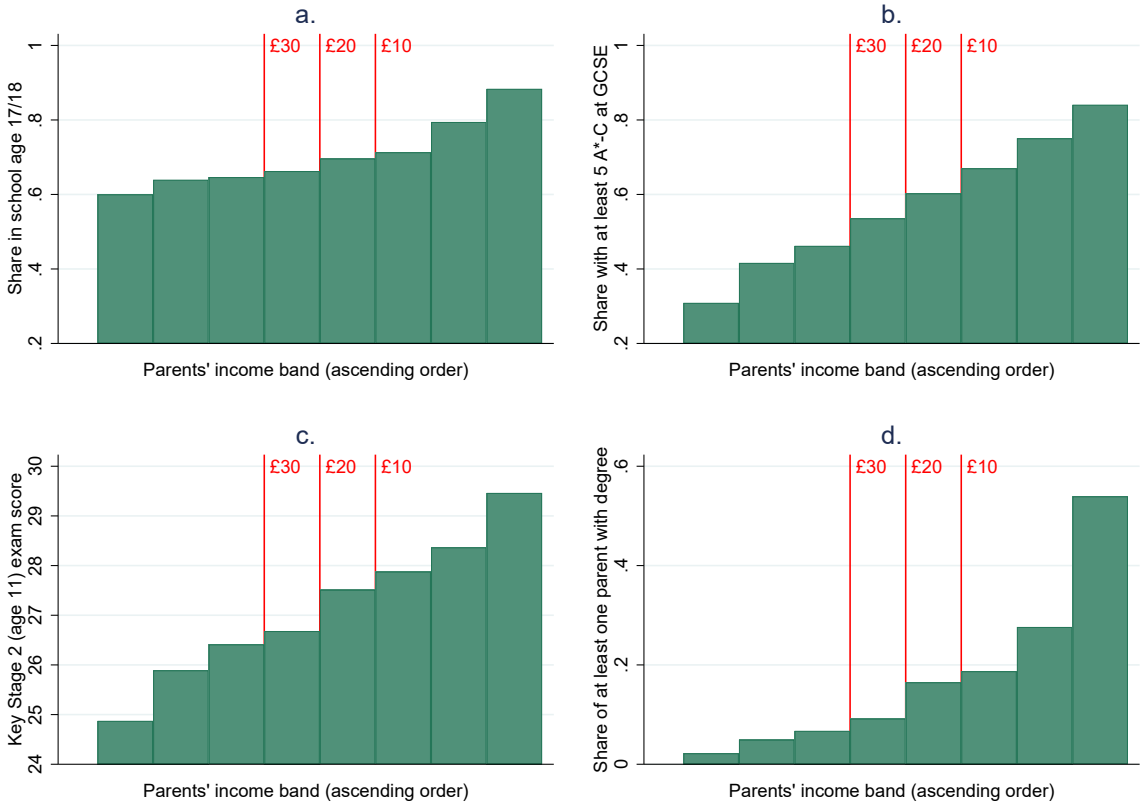
Figure 8 shows transitions between activities by EMA eligibility groups. For instance, the top left hand panel of the figure shows that among individuals in their late teens who were ineligible for the EMA, around 80% of those in education (or training) in the previous period continued in education in the following period, while around 20% transitioned into work. A relatively low share of these individuals transitioned from school to NEET.

Finally, Figure 9 provides some indication of the potential effect of the EMA and other conditional transfers on post-compulsory education participation. Panel *a* shows education participation at age 17/18 by age 16 parental income band. As might be expected, participation is highest for students from the top income group at around 88%, and declines monotonically to around 60% for students from the poorest families.

However, what is notable is that this gradient exhibits a non-linearity: differences between the bottom six parental income groups are much smaller than differences between the top three. Strikingly, the same pattern is not observed in achievement of five A*-C grades at GCSE,

a key metric for age 16 attainment, shown in Panel *b*. Nearly twice as many students from the poorest families are in education at age 17/18 as achieve five A*-C grades at GCSE, whereas for students from the top three parental income groups, the rates are roughly the same. Participation is the outlier here: Panel *c* shows that age 11 (*Key Stage 2*) test scores follow a similar pattern as GCSE attainment. Panel *d* shows that parental education level, which might be another important determinant of child education participation, also falls steeply with income.

Figure 9: Selected education variables by parental income band



Notes. The parental (household) income bands shown are, from left to right: up to £10,399; £10,400 – £15,599; £15,600 – £20,799; £20,800 – £25,999; £26,000 – £31,199; £31,200 – £41,599; £41,600 – £51,999; £52,000 or more. Red lines show the approximate locations of the EMA eligibility thresholds.

One possible explanation for this pattern is that the subsidies provided by the EMA, Child Benefit, and the Child Tax Credit are effective in encouraging students from poor families to participate in post-compulsory education. The observed pattern appears to be consistent with a substantial role for the EMA: the participation rate diverges from the share with at least 5 A*-C at GCSE at roughly the EMA eligibility thresholds (marked in red). This non-linearity provides an important sources of identification for our model. However, it should be noted

that Child Benefit and the Child Tax Credit likely also play an important role, as for most income groups they make up a larger share of the total subsidy.

4.3 Model

Our model is a discrete choice dynamic programming model in the tradition of Keane and Wolpin (1994, 1997). In each of the first ten years after the end of compulsory secondary schooling at age 16, individuals choose between three options: getting more education, seeking employment, or neither. These choices correspond to three employment statuses: in education or training, in employment, or not in employment, education or training (NEET). An individual's choice in one period affects not only her current payoff but also her expected future payoffs. Individuals take their initial endowments and the dynamic effects of their choices into account when deciding between options. Unobserved *ex ante* heterogeneity among individuals is captured by a type variable that is known to the agent but not to the econometrician.

We extend this standard career choice framework in two ways. First, as in Eckstein and Wolpin (1999) and Cameron and Heckman (2001), the future payoff from a year of education in our model depends on whether the individual passes the grade, which is a stochastic function of her initial endowment and past choices. Second, we incorporate a job search into the model, so that individuals who seek work may fail to find a job and become unemployed (NEET). This accounts for the empirical fact that a large share of NEET individuals report being unemployed. The job finding probability is a function of an individual's initial endowment and past employment statuses.

Ten years after the end of compulsory schooling, individuals in our model receive a terminal value, which is intended to reflect their discounted future utility stream at age 26.[26] The terminal value depends on schooling and work experience accumulated over ten years as well as initial endowments. This greatly reduces the computational complexity of solving the model and thus obviates the need for using approximation techniques in estimation. It also means that there is no requirement for our model to capture labour market dynamics beyond age 26.

Let d_{kt} for $k \in \{S, E, N\}$ be indicators, respectively, of whether an individual chooses further schooling or training, seeks employment, or neither in period t.[27] $d_{kt} = 1$ indicates that

[26] A similar approach is taken in Belzil, Maurel and Sidibé (2021).

[27] Individual subscripts are dropped throughout this section to improve readability.

alternative k was chosen in period t, and $d_{kt} = 0$ indicates otherwise. As the alternatives are mutually exclusive, $\sum_{k\in\{S,E,N\}} d_{kt} = 1$. Similarly, let s_{mt} for $m \in \{S, E, N\}$ be indicators, respectively, of whether an individual is in education or training, in employment, or not in employment, education or training (NEET). In what follows, we call $\boldsymbol{d}_t$ an individual's choice, and $\boldsymbol{s}_t$ her employment status. As will become clear, it is further convenient to define an additional discrete vector $\boldsymbol{s}_t^*$, which represents the 'future-relevant' employment status vector. The difference between $\boldsymbol{s}_t$ and $\boldsymbol{s}_t^*$ is merely that individuals who fail the grade are recorded as NEET in $\boldsymbol{s}_t^*$, which reflects the contribution of their present status to their future prospects.

Per-period utility for each individual at time t is specified as logarithmic in consumption with additively separable components. Let U_{mt} denote the per-period utility flow if $s_{mt} = 1$. The per-period utility flows for $m \in \{S, E, N\}$ are:

$$U_{St} = \log C_{St} + \psi_0^S + \psi_1^S A_{16} + \psi_2^S H - \gamma^r s_{S,t-1}^* + \epsilon_t \tag{28}$$

$$U_{Et} = \log C_{Et} - \gamma^{js} I(s_{E,t-1}^* = 0) \tag{29}$$

$$U_{Nt} = \log C_{Nt} + \psi_0^N + \psi_1^N H - \gamma^{js} d_{Et} + \eta_t \tag{30}$$

C_{mt} is consumption if $s_{mt} = 1$ for $m \in \{S, E, N\}$. $I(.)$ is the indicator function. A_{16} is an indicator for whether an individual achieved at least 5 A*-C at GCSE. H is an unobserved binary type. γ^r is a cost that individuals incur if they return to school having been away in the previous period, or if they were in school in the previous period, but did not pass the grade. γ^{js} is a job search cost parameter. ϵ_t and η_t are stochastic shocks known to the individual at the beginning of period t, but not before.

An individual with employment status $s_{mt} = 1$ faces the budget constraint

$$C_{mt} = Y_{mt} + T_{mt} \tag{31}$$

where Y_{mt} is the agent's own income and T_{mt} is a transfer she receives from her parents. The agent's own income is given by

$$Y_{mt} = s_{St} B_t^{ema} + s_{Et} w_t - \tau s_{St} I(S_{t-1} \geq 2) \tag{32}$$

where B_t^{ema} is the value of the EMA cash transfer an individual receives if she is in education

or training in period t and w_t is the individual's wage if she is employed. τ represents tuition fees and other costs of advanced education beyond secondary school and S_t is the number of successfully completed years of education at the end of period t. Neither young people or their parents can borrow or save.[28]

The wage equation is

$$\log w_t = \psi_0^w + \psi_1^w \log Y^P + \psi_2^w F + \psi_3^w A_{11} + \psi_4^w A_{16} + \beta_1 X_{t-1} + \beta_2 S_{t-1} + \omega_t. \tag{33}$$

where Y^P is parental income, F is an indicator for whether an individual is female, and A_{11} is age 11 school attainment as measured by the total Key Stage 2 SAT score. X_{t-1} is accumulated work experience at the end of period $t-1$. ω_t is a stochastic shock known to the individual at the beginning of period t, but not before. The laws of motion for work experience and successfully completed education are

$$S_{t+1} = S_t + p_t s_{St} = S_t + s_{St}^* \tag{34}$$

and

$$X_{t+1} = X_t + s_{Et} = X_t + s_{Et}^*. \tag{35}$$

The vector of stochastic shocks $\boldsymbol{e}_t = (\epsilon_t, \omega_t, \eta_t)'$ is distributed multivariate normal with mean zero and variance Σ. Shocks are thus allowed to be contemporaneously correlated but are assumed to be serially independent.

The value of the transfer an agent receives from her parents is

$$T_{mt} = \max\left[0, \frac{\chi}{1+\chi}\left(Y^P + B_{mt}^P\right) - \frac{Y_{mt}}{1+\chi}\right] \tag{36}$$

B_{mt}^P is the sum of benefits an individual's parents receive if $s_{mt} = 1$ for $m \in \{S, E, N\}$.[29] A simple household bargaining model that rationalizes this functional form for the transfer T_{mt} is presented in Appendix B.

[28] This assumption is motivated by the empirical lack of credit available to low-income households. Young people from such households, who are our focus in this paper, are especially likely to be borrowing-constrained.

[29] More specifically, B_{St}^P is the sum of Child Benefit and Child Tax Credit, and $B_{mt}^P = 0$ for $m \in \{E, N\}$. For details on how Child Benefit and Child Tax Credit were determined, see Appendix A.

Grade passing. Using a similar approach to Eckstein and Wolpin (1999), in a given year of schooling, agents fail to pass a grade with probability $1 - P(p_t = 1)$, where p_t is an indicator such that $p_t = 1$ if a student passes a grade and $p_t = 0$ otherwise. Failing a grade has no current period effects — an individual accrues all the current period benefits (and costs) associated with schooling as normal — but future periods are affected, as accumulated schooling does not increase. This affects future rewards from work and school, job offer probabilities and an individual's terminal value.

p_t is a stochastic function of age 16 attainment A_{16} and the previous period's 'future-relevant' education status $s^*_{S,t-1}$:

$$p_t = I(\psi^p_0 + \psi^p_1 A_{16} + \pi s^*_{S,t-1} + \xi_t > 0) \tag{37}$$

where ξ_t is distributed standard logistic. Hence $P(p_t = 1)$ is the logistic function

$$P(p_t = 1) = \frac{1}{1 + \exp\left[-(\psi^p_0 + \psi^p_1 A_{16} + \pi s^*_{S,t-1})\right]}. \tag{38}$$

Job search. In order to obtain work in a given period, an individual must look for job. Job search has two possible outcomes: employment ($s_{Et} = 1$) or unemployment/NEET ($s_{Nt} = 1$). If the individual was not in work in period $t-1$, looking for a job is associated with a fixed cost equal to γ^{js} incurred in period t, regardless of the outcome of the search. If the individual was in work in period $t-1$ she must still 'look for a job', but does not incur the job search cost if she stays in work. This takes into account that many workers keep their job across periods but allows for a positive probability of job loss.

Let j_t be an indicator such that an individual receives a job offer if $j_t = 1$ and no offer if $j_t = 0$. Let $P(j_t = 1)$ be an individual's *ex ante* probability that she will receive a job offer in period t. j_t is given by

$$j_t = I(\boldsymbol{\psi}^j \boldsymbol{z}^j + \lambda_1 S_{t-1} + \lambda_2 X_{t-1} + \lambda_3 s^*_{S,t-1} + \lambda_4 s^*_{E,t-1} + \zeta_t > 0) \tag{39}$$

where ζ_t is distributed standard logistic and

$$\boldsymbol{\psi}^j \boldsymbol{z}^j = \psi^j_0 + \psi^j_1 \log Y^P + \psi^j_2 F + \psi^j_3 A_{11} + \psi^j_4 A_{16}.$$

Hence the probability λ_t is the following logistic function of initial conditions, successfully completed years of education, accumulated work experience, and (successful) education and employment in the previous period:

$$P(j_t = 1) = \frac{1}{1 + \exp\left[-\left(\boldsymbol{\psi}^j \boldsymbol{z}^j + \lambda_1 S_{t-1} + \lambda_2 X_{t-1} + \lambda_3 s^*_{S,t-1} + \lambda_4 s^*_{E,t-1}\right)\right]}. \tag{40}$$

Initial conditions. Individuals enter the model at the beginning of the first year of non-compulsory schooling with a set of initial conditions. These are: Parental income, Y^P; a dummy variable indicating whether at least one parent has received advanced education, S^P; A dummy variable indicating whether an individual is female, F; a dummy variable indicating whether an individual is the eldest child in the family, E; the score on a standardized exam an individual achieved at age 11 (*Key Stage 2* SATs), A_{11}; a dummy indicating whether the student achieved a 'good pass' (five A*-C grades) on their GCSE (age 16) exams, A_{16}; and an unobservable binary type heterogeneity, $H \in \{0, 1\}$. As set out in the preceding subsections, different initial conditions enter different model equations.

Initial conditions vary across individuals but are fixed over time. This implies that parental income — and hence EMA eligibility — is held constant over time. This simplifies the model; in fact, EMA eligibility was in many cases also determined upon entry to post-compulsory schooling, rather than at the start of each year.[30]

The unobserved type heterogeneity, H, is intended to reflect exogenous taste variation that is unobservable in the data. H affects the utility value of education or training, and the utility value of neither being in work nor in education. The probability of $H = 1$ is given by

$$P(H = 1) = \frac{1}{1 + \exp\left[-(\psi^h_0 + \psi^h_1 S^P + \psi^h_2 A_{16})\right]}. \tag{41}$$

The type heterogeneity is assigned according to this probability when simulating from the model.

One complication arising from data access constraints is that the school attainment variables A_{11} and A_{16} are only available within a secure environment, but for computational reasons the model needs to be estimated outside the secure environment. We deal with this prob-

[30] This is not true for individuals who did not go to school at 16, but subsequently returned, since for them parental income and employment at 17 rather than 16 would matter. This is unlikely to substantively affect the results, however, as this proportion of individuals is low, and EMA eligibility at 16 is likely to be highly correlated with EMA eligibility at 17.

lem by estimating two regression models in the secure environment that summarize the correlation structure between the prior attainment variables and other initial conditions. We assume that

$$P(A_{16} = 1) = \frac{1}{1 + \exp\left[-(a_0 + \sum_{i=2}^{8} a_{i-1} I(YB_i^p = 1) + a_8 E + a_9 F + a_{10} S^P)\right]} \tag{42}$$

where YB_i^p is the income band of individual i's parents.[31] The parameters of this model can be estimated using logistic regression. We also estimate the linear regression model

$$A_{11} = b_0 + \sum_{i=2}^{8} b_{i-1} I(YB_i = 1) + b_8 E + b_9 F + b_{10} S^P + b_{11} A_{16} + \sigma \epsilon_i^{11} \tag{43}$$

where ϵ_i^{11} is assumed to be standard normal.

Estimated parameter values for these regression models are given in Appendix E. In a second step, we use these estimated parameters to simulate prior attainment (i.e. A_{16} and A_{11}) for each individual in the data using the estimated parameters. To the extent that these parameters reflect the true correlation structure in the population, the resulting 'synthetic' sets of initial conditions will still be an unbiased sample from the population distribution.

Terminal value functions. The model terminates after $T = 10$ periods with a terminal value that is a nonlinear function of the post-compulsory education and work experience an individual acquired during the ten periods, as well as her age 16 school attainment.

$$V_{T+1} = \alpha_1 S_T + \alpha_2 S_T^2 + \alpha_3 X_T + \alpha_4 X_T^2 + \alpha_5 A_{16} S_T + \alpha_6 A_{16} X_T. \tag{44}$$

The terminal value function is intended to reflect the discounted stream of future utility at age 26.

Dynamic optimization. At any time t, an individual chooses the option that will maximize the expected present value of her discounted lifetime utility. There are three sources of risk in the model: job market risk, grade-passing risk, and the stochastic elements of the payoff in each employment status.

[31] The eight income bands are the same as in Figure 9: up to £10,399; £10,400 – £15,599; £15,600 – £20,799; £20,800 – £25,999; £26,000 – £31,199; £31,200 – £41,599; £41,600 – £51,999; and £52,000 or more.

Define Ω_t as the information set at the beginning of period t

$$\Omega_t = \{Y^P, S^P, F, E, A_{11}, A_{16}, H, S_{t-1}, X_{t-1}, \boldsymbol{s}^*_{t-1}, \boldsymbol{e}_t\}. \tag{45}$$

Future utility at the optimal choice is given by the value function

$$V_t(\Omega_t) = \max_{\boldsymbol{d}_t} \mathrm{E}\left[\left[\sum_{\tau=t}^{T} \sum_{m\in\{S,E,N\}} \delta^{\tau-t} U_{m\tau}(\Omega_\tau) P(s_{m\tau})\right] + \delta^{T-t+1} V_{T+1}(S_T, X_T, A_{16}) \Big| \Omega_t, \boldsymbol{d}_t\right]. \tag{46}$$

The expectation is taken over the joint distribution of future stochastic shocks to preferences and productivity $\{\boldsymbol{e}_\tau\}_{\tau=t+1}^T$, and the shocks associated with the job market $\{\zeta_\tau\}_{\tau=t}^T$ and grade passing $\{\xi_\tau\}_{\tau=t}^T$. For $t = 1, \ldots, T$, per-period utility is given by equations (28),(29), and (30). At the end of period T, agents receive the terminal value V_{T+1}.

The maximization problem can be recast as a dynamic program by writing $V_t(\Omega_t)$ as the maximization over alternative-specific value functions $V_{kt}(\Omega_t, \boldsymbol{d}_t)$ for $k \in \{S, E, N\}$:

$$V_t(\Omega_t) = \max[V_{St}(\Omega_t), \mathrm{E}(V_{Et}(\Omega_t)), V_{Nt}(\Omega_t)] \tag{47}$$

where the expectations operator represents job market risk arising from ξ_t. These in turn can be written in terms of status-specific value functions $\tilde{V}_m(\Omega_t, t)$ for $m \in \{S, E, N\}$. The alternative-specific value functions are

$$V_{St}(\Omega_t, \boldsymbol{d}_t) = \tilde{V}_{St}(\Omega_t, \boldsymbol{d}_t), \tag{48}$$

$$V_{Et}(\Omega_t, \boldsymbol{d}_t) = P(j_t = 1)\tilde{V}_{Et}(\Omega_t, \boldsymbol{d}_t) + [1 - P(j_t = 1)]\tilde{V}_{Nt}(\Omega_t, \boldsymbol{d}_t), \tag{49}$$

and

$$V_{Nt}(\Omega_t, \boldsymbol{d}_t) = \tilde{V}_{Nt}(\Omega_t, \boldsymbol{d}_t). \tag{50}$$

The status-specific value functions all take the Bellman form

$$\tilde{V}_{mt}(\Omega_t, \boldsymbol{d}_t) = U_{mt}(\Omega_t, \boldsymbol{d}_t) + \delta\, \mathrm{E}[V_{t+1}(\Omega_{t+1})|s_{mt} = 1]. \tag{51}$$

The expectation in equation (51) is taken over the random elements of Ω_{t+1} given $\boldsymbol{s}_t$: S_t and $\boldsymbol{s}^*_{t-1}$, which are functions of ξ_t, and $\boldsymbol{\epsilon}_{t+1}$. This expectation needs to be evaluated by Monte Carlo integration. The model can then be solved by backward recursion. Given the extensive *ex ante* heterogeneity between individuals, no two individuals have the same initial conditions,

so the model needs to be solved separately for each individual.[32]

4.4 Estimation

The model is estimated using our extended version of the Generalized Indirect Inference (GII) method of Bruins et al. (2018). Maximum Likelihood estimation is infeasible, because only selected moments from the administrative schools data that we use can be exported from a secure environment. GII with one-step-ahead forecasts is also impossible, because the model features unobserved heterogeneity in various forms: most obviously, the type heterogeneity H is unobserved, but so are the indicator whether an agent passes the grade p_{jt} and — as they need to be simulated based on regression coefficients exported form a secure environment — the prior attainment parameters A_{11} and A_{16}.

Our auxiliary model has a total of 259 parameters that capture the relevant patterns in the data. Most are direct moments of the data. These are: the status shares by EMA eligibility group (£30, £20, £10, ineligible) in each period; the average log earnings of those eligible and those ineligible for EMA in each period; the transition matrices for the transitions between different activities, again split by time period and EMA eligibility group; the share of NEET individuals reporting to be unemployed in each period in each EMA eligibility group; and the share of individuals in each EMA eligibility group reporting that they live with their parents or receive a financial transfer from their parents at age 26 (both of which we interpret as a 'transfer' in the model). Furthermore, we include the parameters of the wage equation (33) and of a linear probability model of (40), both estimated by Ordinary Least Squares on the selected sample in work and looking for a job, respectively.[33] The number of parameters in the structural model is 45. All except the discount rate δ are estimated using our GII procedure. As is common in the literature, δ is calibrated as $\delta = 0.95$ because it is not well-identified.

Smoothing in our model is more complex than in the generic model outlined in section 2, as it includes not one but five discrete endogenous variables: each period's choice vector $\boldsymbol{d}_t$, each period's employment status vector $\boldsymbol{s}_t$, each period's 'future-relevant' employment status vector $\boldsymbol{s}_t^*$, the indicator for whether an agent in school passes the grade p_t, and the indicator for whether an individual receives a job offer j_t. In addition, it features the discrete unobserved heterogeneity term H. Let $\mathfrak{s}_{t-1}^*$ be shorthand for $\{\{\boldsymbol{s}_\tau^*\}_{\tau=1}^{t-1}, H\}$ and Ω_t^* denote the exogenous

[32] We use $1,000$ simple random draws to evaluate each expectation.

[33] Both regressions were estimated in the secure environment. For details on the construction of the model moments and their smooth analogues, see Appendix C.

elements of the information set Ω_t. Then $\mathfrak{s}^*_{t-1}$ and Ω^*_t together uniquely determine Ω_t.

As outlined in section 2.3, we obtain a smooth objective surface through a weighting procedure that smoothly reweights between the path that each simulated individual takes through the model and any neighbouring paths. In this model, weights depend on the latent utility payoffs from different options, but also on the latent job finding and grade passing draws. For a succinct statement of the weights, it is again useful to define the vector $\boldsymbol{v}_{kt}(\mathfrak{s}^*_{t-1}, \Omega^*_t)$, each element of which is given by $u_{kt}(\mathfrak{s}^*_{t-1}, \Omega^*_t) - u_{jt}(\mathfrak{s}^*_{t-1}, \Omega^*_t)$, where j ranges over all $j \in \{S, E, N\}$ such that $j \neq k$.

In the generic model of section 2.3, choices, employment statuses, and future-relevant employment statuses are all the same. In contrast, employment statuses come apart from choices in this model because individuals do not always get what they choose: people who choose to seek work may end up as NEET if they do not find work. Furthermore, as discussed above, future-relevant employment statuses are not the same as employment statuses, because those that are in education may fail to pass the grade and thus count as NEET for the purposes of calculating future payoffs.

It is therefore helpful to distinguish between 'choice weights', 'status weights' and 'future-relevant status weights' in this model. Choice weights are relevant for obtaining smooth moments that depended on individuals' choices; in our model, they are only used for the calculation of smooth unemployment shares. Status weights are relevant to the calculation of all auxiliary parameters in our model, as all depend on individuals' statuses. Future-relevant status weights are needed primarily to calculate the relevant weights for the following period. Details on how we construct smoothed auxiliary parameters are given in Appendix C.

Choice weights are given by

$$W^c_{kt}(\mathfrak{s}^*_{t-1}, \Omega^*_t; \boldsymbol{\rho}) = \mathcal{K}_2 \left[\boldsymbol{v}_{kt}(\mathfrak{s}^*_{t-1}, \Omega^*_t) / \rho_2 \right] W_t(\mathfrak{s}^*_{t-1}, \Omega^*_{t-1}; \boldsymbol{\rho}) \tag{52}$$

where $\mathcal{K}_2$ is the multivariate kernel given by (10), $\boldsymbol{\rho} = (\rho_1, \rho_2, \phi)$ is a vector of smoothing parameters, and $W_t(\mathfrak{s}^*_{t-1}, H, \Omega^*_t; \boldsymbol{\rho})$ reflects the weight of the state at the beginning of the period.[34]

[34] Choice weights are analogous to the summands in (7): smoothed choices could be obtained as

$$\widetilde{\widetilde{d_{kt}(\rho)}} = \sum_{\mathfrak{s}^*_{t-1} \in S_{t-1}} W^c_{kt}(\mathfrak{s}^*_{t-1}, \Omega^*_t; \boldsymbol{\rho}).$$

We set $\rho_1 = 0.03$, $\rho_2 = 0.01$ and $\phi = 2,000$.

Status weights are given by

$$W^s_{St}(\mathfrak{s}^*_{t-1}, \Omega^*_t; \boldsymbol{\rho}) = W^c_{St}(\mathfrak{s}^*_{t-1}, \Omega^*_t; \boldsymbol{\rho}) \tag{53}$$

$$W^s_{Et}(\mathfrak{s}^*_{t-1}, \Omega^*_t; \boldsymbol{\rho}) = W^j_t(\mathfrak{s}^*_{t-1}, \Omega^*_t) W^c_{Et}(\mathfrak{s}^*_{t-1}, \Omega^*_t; \boldsymbol{\rho}) \tag{54}$$

$$W^s_{Nt}(\mathfrak{s}^*_{t-1}, \Omega^*_t; \boldsymbol{\rho}) = W^c_{Nt}(\mathfrak{s}^*_{t-1}, \Omega^*_t; \boldsymbol{\rho}) + \left(1 - W^j_t(\mathfrak{s}^*_{t-1}, \Omega^*_t; \rho_1)\right) W^c_{Et}(\mathfrak{s}^*_{t-1}, \Omega^*_t; \boldsymbol{\rho}) \tag{55}$$

where $W^j_t(\mathfrak{s}^*_{t-1}, \Omega^*_t; \rho_1) \in [0, 1]$ denotes the weight factor of finding a job. *Future-relevant status weights* are given by

$$W^{s*}_{St}(\mathfrak{s}^*_{t-1}, \Omega^*_t; \boldsymbol{\rho}) = W^p_t(\mathfrak{s}^*_{t-1}, \Omega^*_t; \rho_1) W^s_{St}(\mathfrak{s}^*_{t-1}, \Omega^*_t; \boldsymbol{\rho}) \tag{56}$$

$$W^{s*}_{Et}(\mathfrak{s}^*_{t-1}, \Omega^*_t; \boldsymbol{\rho}) = W^s_{Et}(\mathfrak{s}^*_{t-1}, \Omega^*_t; \boldsymbol{\rho}) \tag{57}$$

$$W^{s*}_{Nt}(\mathfrak{s}^*_{t-1}, \Omega^*_t; \boldsymbol{\rho}) = W^s_{Nt}(\mathfrak{s}^*_{t-1}, \Omega^*_t; \boldsymbol{\rho}) + \left(1 - W^p_t(\mathfrak{s}^*_{t-1}, \Omega^*_t; \rho_1)\right) W^c_{St}(\mathfrak{s}^*_{t-1}, \Omega^*_t; \boldsymbol{\rho}). \tag{58}$$

where $W^p_t(\mathfrak{s}^*_{t-1}, \Omega^*_t; \rho_1) \in [0, 1]$ is the weight factor associated with passing or not passing a grade. Completing the recursive definition, $W^{s*}_{kt}(\mathfrak{s}^*_{t-1}, \Omega^*_t; \boldsymbol{\rho})$ can alternatively be written as

$$W_{t+1}(\mathfrak{s}^*_t, \Omega^*_t; \boldsymbol{\rho}) = \sum_{k \in \{S,E,N\}} s^*_{kt} W^{s*}_{kt}(\mathfrak{s}^*_{t-1}, \Omega^*_t; \boldsymbol{\rho}). \tag{59}$$

$W(\mathfrak{s}^*_0, \Omega^*_0; \boldsymbol{\rho}) = W(H, \Omega^*_0; \rho_1)$ is the weight associated with H, where Ω^*_0 is the time-invariant subset of Ω_t.[35]

For each binary discrete endogenous variable $x \in \{p, j\}$, the associated weights are

$$W^x_t(\mathfrak{s}^*_{t-1}, \Omega^*_t; \rho_1) = \mathcal{K}_1(x^*_t(\mathfrak{s}^*_{t-1}, \Omega^*_t)/\rho_1) \tag{60}$$

where $\mathcal{K}_1$ is a univariate kernel given in (9). $x^*_t(\mathfrak{s}^*_{t-1}, \Omega^*_t)$ is the relevant latent value and ρ_1 is a smoothing parameter. For the weight associated with the unobserved heterogeneity term H, we similarly set

$$W(H = 1, \Omega^*_0; \rho_1) = \mathcal{K}_1(H^*(\Omega^*_0)/\rho_1) \tag{61}$$

[35] In analogy to (8), it would be possible to normalize the weights here by dividing by $\sum_{k \in \{S,E,N\}} W^{s*}_{kt}(\mathfrak{s}^*_{t-1}, \Omega^*_t; \boldsymbol{\rho})$. We leave out this step here, because the relatively moderate size of the state space allows us to set $\phi = 2000$, so $\sum_{k \in \{S,E,N\}} W^{s*}_{kt}(\mathfrak{s}^*_{t-1}, \Omega^*_t; \boldsymbol{\rho})$ is very close to one in all cases.

and $W(H = 0, \Omega_0^*; \rho_1) = 1 - W(H = 1, \Omega_0^*; \rho_1)$.[36]

We minimize the criterion function $C(\theta)$ given by (4) using a version of the Levenberg-Marquardt optimization algorithm with an ellipsoidal trust region.[37] This algorithm requires information about the Jacobian of the least squares residuals $\boldsymbol{r}(\boldsymbol{\theta}) = \mathcal{W}^{1/2}(\tilde{\boldsymbol{\varphi}}(\boldsymbol{\theta}, \boldsymbol{\rho}) - \hat{\boldsymbol{\varphi}})$, but no information about second derivatives; the Hessian of $C(\theta)$ is approximated based on the Jacobian of $\boldsymbol{r}(\boldsymbol{\theta})$.[38] The Levenberg-Marquardt Algorithm is well-suited for our problem for three reasons. First, it only requires the objective function to be continuously differentiable and not twice differentiable, which would not be true of our objective function. Second, it can be efficiently parallelized, as derivative evaluations are independent. Third, the approximation of the Hessian based on the Jacobian of the residuals is highly effective, which leads to very fast convergence.[39]

4.5 Identification

As is typical in complex dynamic discrete choice models, parameter identification is not straightforward in this model. Various exclusion restrictions and functional form assumptions together identify the parameters of our model. Here we give an informal overview of which moments are primarily affected by which model parameters and are therefore important for pinning down the parameter values.

The parameters of the status-specific flow utility functions for education and NEET status, (28) and (29), as well as those of the grade passing process (38), are primarily pinned down by the status shares in each EMA eligibility group and the transition probabilities between them. For instance, a large coefficient on age 16 attainment (A_{16}) in the flow utility function for education will lead to a higher propensity to stay in education for those ineligible for the EMA, who tend to have higher school attainment. Large coefficients on the unobserved heterogeneity would lead to more persistence in employment statuses.

The parameters of the wage function are primarily identified by the parameters of the reduced-form wage regression included in the auxiliary model. A number of exclusion re-

[36] It would also have been feasible to directly weight by the conditional probability given by (41). Our procedure saves computation time at some cost in terms of simulation precision.

[37] We only simulate $M = 1$ data set from the model. As our sample is relatively large and the 'effective sample size' is likely even higher given substantial smoothing, the costs in terms of statistical efficiency are likely to be minor.

[38] We use central differences to approximate the Jacobian of $\boldsymbol{r}(\boldsymbol{\theta})$ numerically. Our problem is appropriately rescaled to maximize the accuracy of this approximation.

[39] For details on the Levenberg-Marquardt algorithm, see e.g. Nocedal and Wright (2006).

strictions help capture potential selection effects. For instance, successful education and employment in the previous period, $s^*_{S,t-1}$ and $s^*_{E,t-1}$, can affect choices through the grade passing and job finding probability, but do not directly affect wages. Similarly, parental education S^P can affect flow utilities and thus choices through the unobserved heterogeneity, but does not directly affect wages.

The parameters of the job finding process are identified mainly through the parameters of the reduced-form job-finding regression included in the auxiliary model, in combination with the unemployment shares by age and EMA eligibility group. The exclusion of parental education S^P from the job finding equation further aids identification.

Among the remaining model parameters, the household bargaining parameter χ is primarily identified by the share of parents making positive transfers to their child when the child is 26; the tuition fee τ is largely pinned down by the observed shares of individuals in education or training beyond age 18 in each EMA eligibility group; the cost of re-entry into education γ^r directly affects the likelihood of transitions back into education or training; and the job search cost γ^{js} is mostly pinned down by the observed unemployment shares.

The parameters governing the assignment of the unobserved heterogeneity in (41), the parameters of the terminal value function (44) and the parameters governing the structural shocks in the model are linked to model moments in a more indirect way. To some extent they pick up residuals not otherwise captured in the model and may thus be less likely to reflect the structure of the true model. As a result, additional care is required in their interpretation.

4.6 Parameter Estimates

Table 3 shows the estimated parameters on state variables that enter linearly in the main model equations. Each column stands for one of the main equations in the model. Empty fields indicate that the state variable in this row does not enter into the equation in this column. Asymptotic standard errors are given in parentheses. All other parameter estimates are presented in Appendix D.

The first column lists the parameters governing utility from education. The coefficient on A_{16}, an indicator for at least 5 A*-C at GCSE, is very large and positive, indicating that prior attainment is an important determinant of post-compulsory education enrolment. The effect of the unobserved type H is nearly equally strong, indicating that *ceteris paribus*, $H = 1$ types are much more likely to enrol in post-compulsory education than $H = 0$ types. The intercept

is large and negative, suggesting that few individuals without 5 A*-C and with $H = 0$ will choose to continue their education after age 16.

Table 3: Estimated parameters: main equations

	Utility (educ.)	**Utility (NEET)**	**Wage**	**Grade passing**	**Job finding**	**Heterogeneity**	**Term. value**
Equation	(28)	(30)	(33)	(38)	(40)	(41)	(44)
Dep. var	U_{st}	U_{nt}	$\log w_t$	p_t	j_t	H	V_{T+1}
Intercept	-7.170	-0.997	7.336	-0.118	-7.847	0.686	
	(0.802)	(0.242)	(0.093)	(0.052)	(0.327)	(0.040)	
$\log Y^P$			0.110		0.330		
			(0.008)		(0.028)		
S^P						1.576	
						(0.118)	
F			-0.239		0.437		
			(0.015)		(0.056)		
A_{11}			0.011		0.171		
			(0.003)		(0.011)		
A_{16}	13.360		0.103	0.323	0.271	-0.964	
	(1.371)		(0.025)	(0.059)	(0.071)	(0.051)	
H	10.626	-4.410					
	(1.102)	(0.472)					
S_{t-1}			0.185		0.019		0.655
			(0.011)		(0.021)		(1.265)
S_{t-1}^2							-4.278
							(0.460)
X_{t-1}			0.125		0.196		-3.325
			(0.011)		(0.023)		(0.462)
X_{t-1}^2			-0.008				-0.015
			(0.001)				(0.017)
$A_{16}S_{t-1}$							-0.388
							(1.376)
$A_{16}X_{t-1}$							4.010
							(0.416)
$s_{S,t-1}^*$				0.334	1.836		
				(0.046)	(0.112)		
$s_{E,t-1}^*$					2.615		
					(0.082)		

Notes. The table shows estimated parameters on state variables that enter linearly in different model equations. Each column stands for one equation/dependent variable. Empty fields indicate that the state variable in this row does not enter into the equation in this column. Asymptotic standard errors are given in parentheses.

The second column lists the parameters governing utility from NEET status. The negative intercept term suggests that consistent with the evidence presented in Figure 6, few individuals will voluntarily choose NEET. The unobserved heterogeneity term is large and negative, suggesting that $H = 1$ individuals will seek to avoid NEET status at a large cost.

The wage parameters listed in the third column are precisely estimated, as they are tightly tied down by the parameters of the reduced form wage regression included in the auxiliary model. All coefficients have the expected signs. All else equal, a 1% increase in parental household income increases child earnings by 0.11%. Women with similar background characteris-

tics earn around 21% less than men. Both age 11 and age 16 school attainment (A_{11} and A_{16}) have a substantial positive effect on earnings. The return to successfully completed schooling is large at around 18% per year. This is around twice as large as standard estimates of the return to schooling from the literature, which is consistent with the fact that the structural parameter reflects the return to *successful* schooling in the model, and in our model, nearly half of all years of post-compulsory schooling are not completed successfully (unsuccessful years of schooling do not increase earnings). The return to work experience X_{t-1} is substantial in the initial year at around 12% but declines rapidly with further years of experience.

The fourth column indicates that *ceteris paribus*, as is intuitive, those with at least 5 A*-C at GCSE have a higher pass probability than those without. Similarly, continuing students who passed in the previous year have a higher pass probability than those re-entering education or who did not pass the grade in the previous year.

The coefficients on state variables in the job finding equation, listed in fifth column, also have the expected sign. The job finding probability rises with parental income ($\log Y^P$), prior attainment (A_{11} and A_{16}), successfully completed education (S_{t-1}) and work experience (X_{t-1}). It is also substantially higher for those coming from successful education ($s^*_{S,t-1}$) or employment ($s^*_{E,t-1}$) than for those who were NEET in the previous period. Interestingly, *ceteris paribus*, women have a higher job finding probability than men. Again, tight standard errors on the parameters of the job finding equation seem to point to the effectiveness of including OLS estimates from a linear probability model of job finding in the auxiliary parameter vector.

The sixth column lists the parameters governing assignment to each unobserved type. Those with highly educated parents are more likely to be type $H = 1$, whereas those who achieve at least 5 A*-C at GCSE are less likely to be type $H = 1$. The large positive coefficient on the type dummy listed in the first and second columns thus likely reflects the fact that children of highly educated parents are likely to obtain post-compulsory education and unlikely to choose NEET status even if their prior attainment is low.

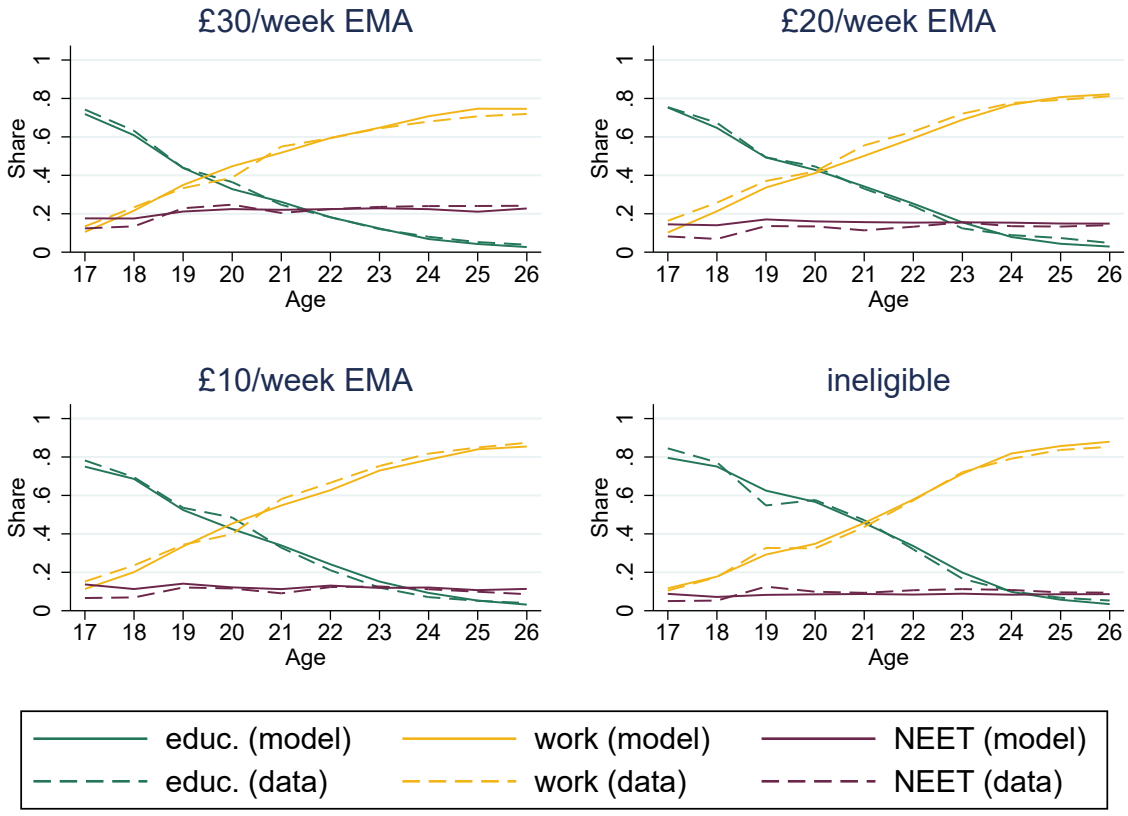
The last column relates to the terminal value function (44). The parameters do not seem to reflect plausible values for the future value of schooling and work experience. In particular, the implied future value of successfully completed education is negative, which is implausible. Various factors may account for this. First, young people may systematically underestimate the economic value of post-compulsory education. Second, the assumed discount factor of $\delta = 0.95$ may be higher than young people's true discount factor. Third, young people's deci-

sions may be imperfectly captured by our model — in particular, they may not be dynamically optimal.

4.7 Model Fit

This section examines the fit of the model to key moments of the data, all of which are targeted in estimation. These are: employment status distributions by age, log annual earnings by age, and transitions between employment statuses by age. Status distributions by age are split by EMA eligibility group; log annual earnings by age and transitions between statuses are split between eligible and ineligible individuals.

Figure 10: Choice distribution by age and EMA eligibility

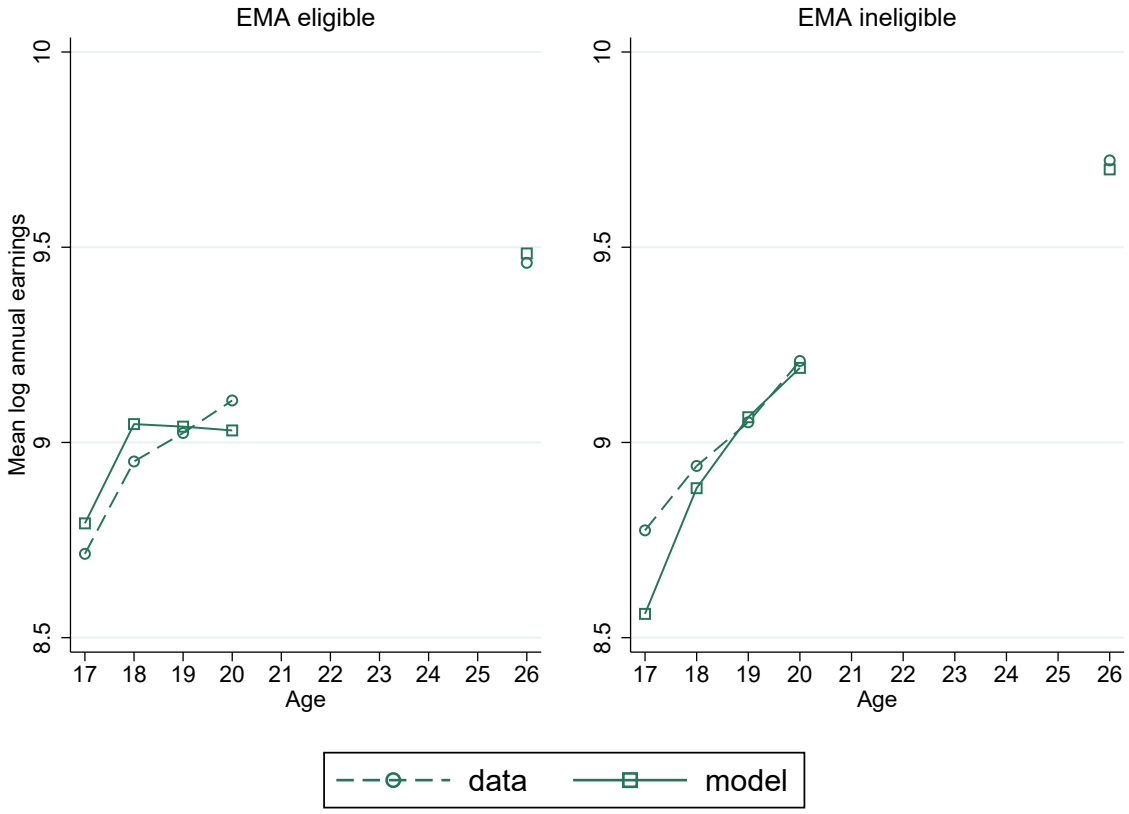


Notes. 'Age' is the young person's age at the end of each period.

Figure 10 presents the model fit for the employment status distribution by age and EMA eligibility group. Dashed lines indicate data values and solid lines model results. The model captures the broad patterns in the data well: The share in education gradually falls with age, while the share in employment rises. It also captures the substantially higher share of young

people in education for young people that are ineligible for the EMA, and the higher NEET share for young people who are eligible for the full EMA amount.

Figure 11: Earnings by Age



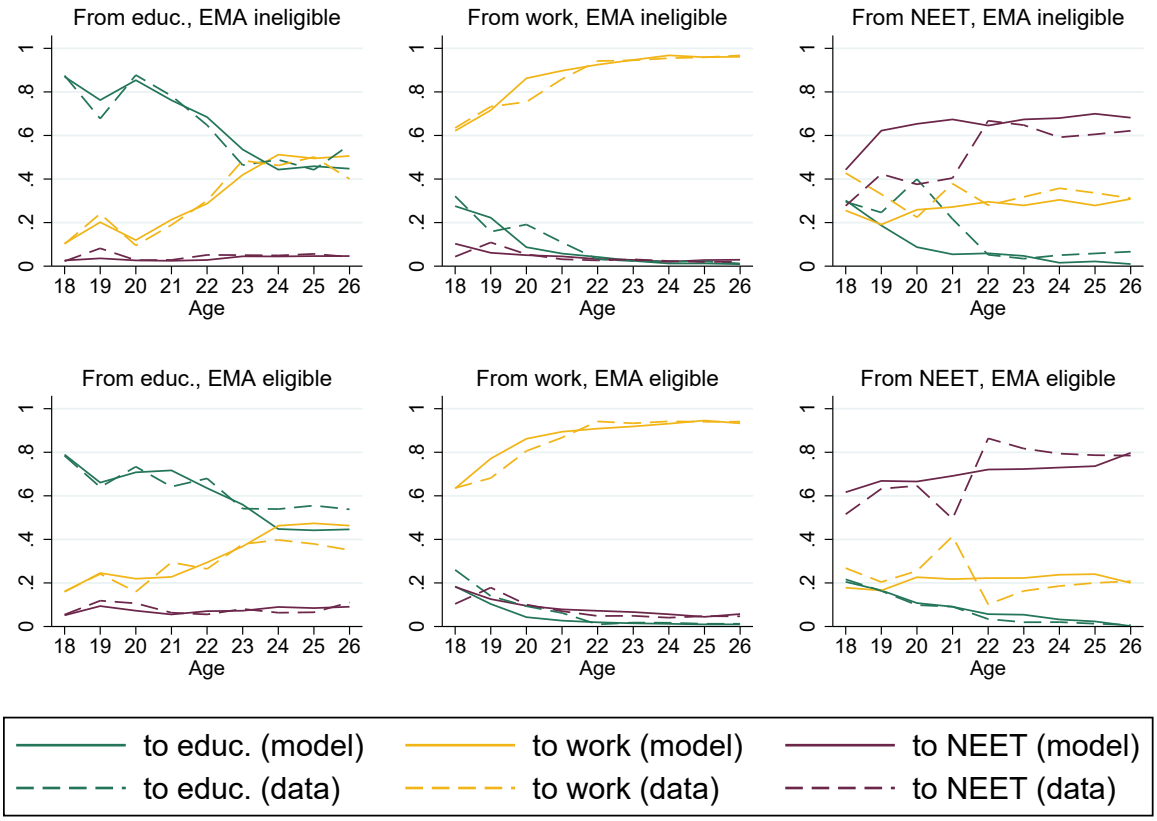
Notes. Mean log earnings for individuals whose main activity is work. Earnings are in 2006 prices. EMA eligibility is based on parental income at age 16. 'Age' is the young person's age at the end of each period.

Figure 11 shows the success of the model at matching the mean log earnings for EMA-eligible and EMA-ineligible individuals. Earnings for EMA-ineligible students are matched very well except for age 16/17, when only a small share of EMA-ineligible individuals are in work as their main activity. The earnings for EMA-eligible individuals are captured less well, especially in their early careers. The model suggests slightly declining average earnings in the late teens for EMA-eligible students, when in fact average earnings are strictly increasing with age in the data. However, it should be noted that the fit is still fairly close, especially considering that the number of people in work at these ages is relatively small.

Figure 12 shows the fit of the model to data on transitions between activities. The fit is excellent for transitions from education and from work, but worse for transitions from NEET. One reason may be the low number of NEET individuals. Another could be the fact that tran-

sition moments from NEET are not explicitly included in estimation, as they are deterministic functions of other moments (albeit non-linear functions).

Figure 12: Transitions by age



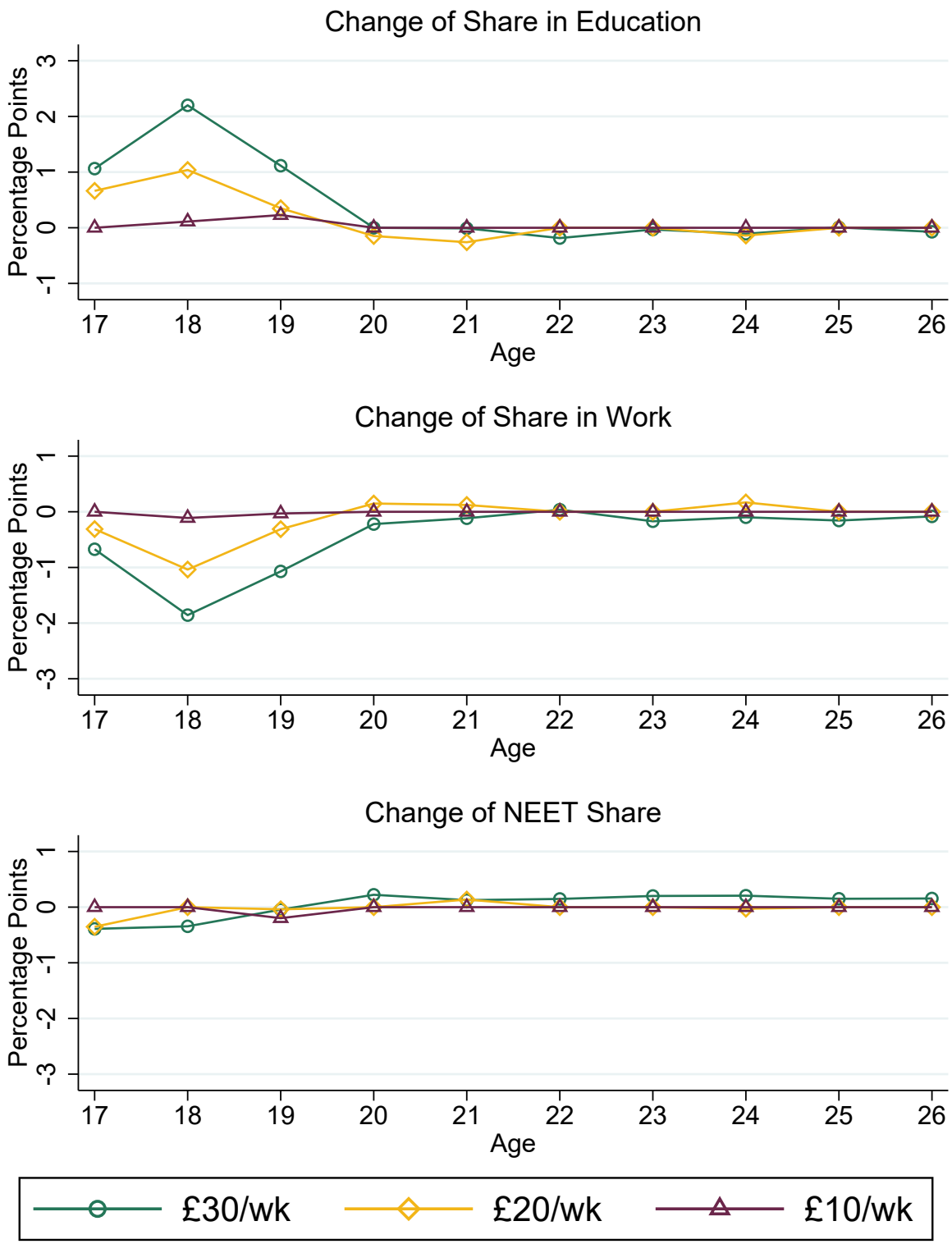
Notes. The diagram shows the share moving into each activity conditional on their activity in the previous year. EMA eligibility is based on parental income at age 16. 'Age' is the young person's age at the end of each period.

4.8 Effects of the EMA

Figure 13 shows the estimated effect of the EMA on employment status shares at different ages against a counterfactual with no grant. The programme modestly increased participation in education or training among eligible students. As the top panel of Figure 13 shows, those eligible for the full grant of £30 per week (£1200 per year) saw by far the largest increases in participation. For this group, the effect was largest at 2.2 percentage points at age 17/18 and somewhat lower at 1.1 percentage points at ages 16/17 and 18/19. There is no persistent effect into individuals' 20s, so the EMA does not seem to encourage more advanced education in our model. Most of the students switching into education switched from employment (middle panel), with a very small impact on the number of NEET individuals. We find much smaller

effects for the partial grant of £20 per week and nearly none for the £10 per week grant.

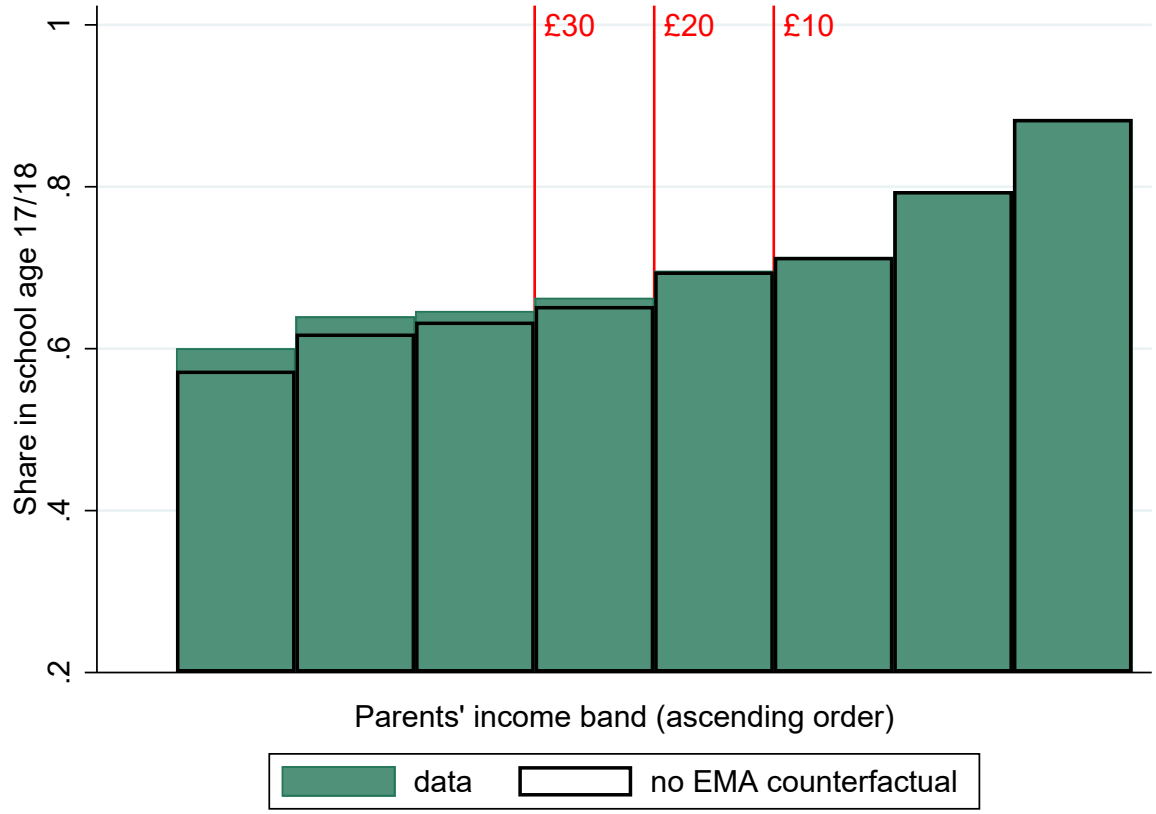
Figure 13: Effect of the EMA on Choice Shares



Notes. 'Age' is the young person's age at the end of each period.

These results imply somewhat smaller effects than those estimated in Dearden et al. (2009) for the pilot programme, but the differences are mostly not statistically significant. They are of similar magnitude as estimates in Chowdry, Dearden and Emmerson (2007) for a later cohort. One reason for the difference from Dearden et al. (2009) might be that they study a cohort that is six years older than ours, so inflation would have eroded the (nominally constant) value of the EMA grant for our cohort compared to theirs by more than 10%. Another reason is that the EMA pilot areas were particularly deprived, so their sample contained a larger share of very poor children, whose behaviour we estimate to be most often affected by the EMA. A third reason is that the overall share of students in post-compulsory education was much lower in earlier cohorts, so more students may have been at the margin.

Figure 14: Age 17/18 effect of the EMA on share in education



Notes. The parental (household) income bands shown are, from left to right: up to £10,399; £10,400 – £15,599; £15,600 – £20,799; £20,800 – £25,999; £26,000 – £31,199; £31,200 – £41,599; £41,600 – £51,999; £52,000 or more. Red lines show the approximate locations of the EMA eligibility thresholds. Counterfactual shares are obtained by subtracting the estimated causal effect of the EMA for each parental income group from the actual shares.

Figure 14 compares the post-compulsory attendance shares at age 17/18 by parental income band to counterfactual attendance shares obtained by subtracting the estimated effect

of the EMA. As the estimated effect of the EMA is largest for the bottom income bands, the socio-economic gradient of the counterfactual shares is steeper than of the actual shares. However, even accounting for the effect of the EMA, post-compulsory attendance shares are still substantially less socio-economically graded among students from lower-income households than measures of age 11 and age 16 school attainment. One factor behind this may be the substantial additional participation subsidies paid to parents in the form of Child Benefit and the Child Tax Credit.

Table 4: Comparing 'long-term compliers' to the broader population

	Whole sample	**EMA eligible**	**Long-term compliers**
Number (unweighted)	8,847	5,118	148
Share (weighted)	1	.560	.016
Background			
Female share	.49	.48	.56
Median household income age 16	28,455	17,927	12,751
Age 11 exam score (*Key Stage 2 SATs*)	27.5	26.6	26.1
Share of good pass (5 A*-C) on age 16 exams (*GCSE*)	.68	.57	.55
Outcomes			
Average years of work experience	4.4	4.4	5.4
Average years of post-compulsory education	3.4	2.9	2.7
Average years of *successful* post-compulsory education	1.9	1.6	1.3
Median earnings (in £)	14,804	13,399	12,324
Mean log earnings	9.59	9.49	9.49
Employment Rate	.82	.78	.87

Notes. All incomes are in 2006 prices. 'Long-term compliers' are individuals for whom the number of successfully completed years of education or years of work experience are different at age 25 with the EMA from what they would have been without the EMA.

As Table 4 shows, we estimate that the EMA had a long-term effect on only around 1.6 percent of all students (3% of EMA-eligible students), which we define as affecting successfully completed years of education or years of work experience by age 25. Compared to other eligible students, these 'long-term compliers' were slightly more likely to be female and from substantially poorer families. Despite their socio-economic background, roughly the same share of individuals within this group achieved at least 5 A*-C at GCSE as the overall EMA-eligible population. Looking at outcomes, 'long-term compliers' completed less education but accumulated substantially more work experience than average. As a result, this group was more likely to be employed at age 26 than both the overall population and the EMA-eligible population. Depending on whether one looks at mean log earnings or median earnings, their earnings were roughly the same or somewhat lower than for the overall EMA-eligible population.

Table 5: 'Long-term compliers' at age 26

	With EMA	**No EMA Counterfactual**
Average years of work experience	5.4	6.5
Average years of post-compulsory education	2.7	1.6
Average years of *successful* post-compulsory education	1.3	.8
Median earnings (in £)	12,324	12,751
Mean log earnings	9.49	9.4
Employment Rate	.87	.89
Share with higher earnings	.54	.46

Notes. Earnings are in 2006 prices. 'Share with higher earnings' excludes individuals who were not in work in either scenario.

Table 5 shows the effect of the EMA only for the subsample of 'long-term compliers'. We estimate that the EMA induced these individuals to enrol for 1.1 years of additional education on average by age 25, of which only around half were successfully completed. This displaced an average of 1.1 years of work experience. Log earnings in this group were on average around 9 log points higher than in a counterfactual scenario without the grant, and employment was nearly unchanged. The net positive effect on average log earnings can be attributed to the fact that 'long-term compliers' acquired a large amount of work experience even with the EMA, and we estimate rapidly diminishing returns to work experience.[40]

However, estimated effects are heterogeneous. We estimate that nearly half of long-term compliers earned less at age 26 as a result of the programme, and *median* earnings are actually slightly lower with the EMA than in the counterfactual scenario. Despite positive earnings effects on average, the additional education participation induced by the EMA leads to lower earnings for a substantial number of individuals who fail to pass the grade.

5 Conclusion

In this paper, we developed a computationally practicable variation of the GII method for use in contexts that exhibit history dependence *and* unobserved heterogeneity. We compared our method to Simulated Maximum Likelihood and conventional GII using *one-step-ahead* simulations in the estimation of the model of Keane and Wolpin (1994), a canonical structural dynamic discrete choice model that can be estimated by all three methods. Our new method roughly matches the other two in terms of both estimation accuracy and performance in a simple policy

[40] From a modelling perspective, the reason that additional education can have substantial pay-offs in expectation can be attributed to the contemporaneous opportunity cost of education, the substantial discount rate, and the lack of a financial market.

experiment.

We then employed the new method to estimate a structural model of the career choices of young people in the UK, which would be prohibitively difficult to estimate using conventional GII. We used this model to assess the effects of the Education Maintenance Allowance (EMA), a conditional cash transfer (CCT) scheme that paid young people up to £3,600 ($4,800) for staying in education or training between the ages of 16 and 19. We found that the policy had a long-run impact on only a small group of very disadvantaged students (less than 2% of all students). For this group, earnings were around 9 percent higher on average at age 26 than in a counterfactual scenario without the grant, but there was substantial heterogeneity of outcomes, with only slightly more than half of affected students obtaining higher earnings. This suggests that CCTs in developed countries with a high base rate of post-compulsory school attendance may only help a very small group of students in the long run, and thus may offer poor value for money unless they are targeted at students from the poorest families.

References

Abaluck, Jason, and Abi Adams-Prassl. 2021. "What Do Consumers Consider Before They Choose? Identification from Asymmetric Demand Responses." *The Quarterly Journal of Economics*. Forthcoming.

Altonji, Joseph G, Anthony A Smith Jr, and Ivan Vidangos. 2013. "Modeling earnings dynamics." *Econometrica*, 81(4): 1395–1454.

Attanasio, Orazio, Erich Battistin, Emla Fitzsimons, and Marcos Vera-Hernandez. 2005. "How effective are conditional cash transfers? Evidence from Colombia." *Institute for Fiscal Studies Briefing Note*, No. 54.

Attanasio, Orazio P, Costas Meghir, and Ana Santiago. 2012. "Education choices in Mexico: Using a structural model and a randomized experiment to evaluate PROGRESA." *The Review of Economic Studies*, 79(1): 37–66.

Bäckman, Olof, Vibeke Jakobsen, Thomas Lorentzen, Eva Österbacka, and Espen Dahl. 2015. "Early school leaving in Scandinavia: Extent and labour market effects." *Journal of European Social Policy*, 25(3): 253–269.

Behrman, Jere R, Jorge Gallardo-Garcia, Susan W Parker, Petra E Todd, and Viviana Vélez-Grajales. 2012. "Are conditional cash transfers effective in urban areas? Evidence from Mexico." *Education Economics*, 20(3): 233–259.

Behrman, Jere R, Susan W Parker, and Petra E Todd. 2011. "Do conditional cash transfers for schooling generate lasting benefits? A five-year followup of PROGRESA/Oportunidades." *Journal of Human Resources*, 46(1): 93–122.

Belzil, Christian, Arnaud Maurel, and Modibo Sidibé. 2021. "Estimating the value of higher education financial aid: Evidence from a field experiment." *Journal of Labor Economics*, 39(2): 361–395.

Berndt, Ernst R, Bronwyn H Hall, Robert E Hall, and Jerry A Hausman. 1974. "Estimation and inference in nonlinear structural models." In *Annals of Economic and Social Measurement, Volume 3, number 4*. 653–665. NBER.

Blundell, Richard, Monica Costa Dias, Costas Meghir, and Jonathan Shaw. 2016. "Female labor supply, human capital, and welfare reform." *Econometrica*, 84(5): 1705–1753.

Blundell, Richard, Monica Costa-Dias, David Goll, and Costas Meghir. 2021. "Wages, Experience, and Training of Women over the Life Cycle." *Journal of Labor Economics*, 39(S1): S275–S315.

Brewer, Mike. 2003. "The new tax credits." *Institute for Fiscal Studies Briefing Note*, No. 54.

Bruins, Marianne, James A Duffy, Michael P Keane, and Anthony A Smith Jr. 2018. "Generalized indirect inference for discrete choice models." *Journal of Econometrics*, 205(1): 177–203.

Brunello, Giorgio, and Maria De Paola. 2014. "The costs of early school leaving in Europe." *IZA Journal of Labor Policy*, 3(1): 22.

Cameron, Stephen V, and James J Heckman. 2001. "The dynamics of educational attainment for black, hispanic, and white males." *Journal of Political Economy*, 109(3): 455–499.

Card, David. 2001. "Estimating the return to schooling: Progress on some persistent econometric problems." *Econometrica*, 69(5): 1127–1160.

Chowdry, Haroon, Lorraine Dearden, and Carl Emmerson. 2007. "Education Maintenance Allowance: Evaluation with Administrative Data." mimeo.

Dearden, Lorraine, and Alexandra Heath. 1996. "Income support and staying in school: What can we learn from Australia's AUSTUDY experiment?" *Fiscal Studies*, 17(4): 1–30.

Dearden, Lorraine, Carl Emmerson, Christine Frayne, and Costas Meghir. 2009. "Conditional cash transfers and school dropout rates." *Journal of Human Resources*, 44(4): 827–857.

De Brauw, Alan, and John Hoddinott. 2011. "Must conditional cash transfer programs be conditioned to be effective? The impact of conditioning transfers on school enrollment in Mexico." *Journal of Development Economics*, 96(2): 359–370.

Eckstein, Zvi, and Kenneth I Wolpin. 1989. "Dynamic labour force participation of married women and endogenous work experience." *The Review of Economic Studies*, 56(3): 375–390.

Eckstein, Zvi, and Kenneth I Wolpin. 1999. "Why youths drop out of high school: The impact of preferences, opportunities, and abilities." *Econometrica*, 67(6): 1295–1339.

Eisenhauer, Philipp, James J Heckman, and Stefano Mosso. 2015. "Estimation of dynamic discrete choice models by maximum likelihood and the simulated method of moments." *International Economic Review*, 56(2): 331–357.

Glewwe, Paul, and Ana Lucia Kassouf. 2012. "The impact of the Bolsa Escola/Familia conditional cash transfer program on enrollment, dropout rates and grade promotion in Brazil." *Journal of Development Economics*, 97(2): 505–517.

Gourieroux, Christian, Alain Monfort, and Eric Renault. 1993. "Indirect inference." *Journal of Applied Econometrics*, 8(S1): S85–S118.

Gregg, Paul, and Emma Tominey. 2005. "The wage scar from male youth unemployment." *Labour Economics*, 12(4): 487–509.

Hajivassiliou, Vassilis. 2000. "Some practical issues in maximum simulated likelihood." *Simulation-based inference in econometrics: methods and applications*, 71–99.

Hajivassiliou, Vassilis, Daniel McFadden, and Paul Ruud. 1996. "Simulation of multivariate normal rectangle probabilities and their derivatives: theoretical and computational results." *Journal of Econometrics*, 72(1-2): 85–134.

Heckman, James J. 1979. "Sample selection bias as a specification error." *Econometrica*, 153–161.

Keane, Michael P, and Kenneth I Wolpin. 1994. "The solution and estimation of discrete choice dynamic programming models by simulation and interpolation: Monte Carlo evidence." *The Review of Economics and Statistics*, 648–672.

Keane, Michael P, and Kenneth I Wolpin. 1997. "The career decisions of young men." *Journal of Political Economy*, 105(3): 473–522.

Kristensen, Dennis, and Bernard Salanié. 2017. "Higher-order properties of approximate estimators." *Journal of Econometrics*, 198(2): 189–208.

Li, Tong, and Bingyu Zhang. 2015. "Affiliation and entry in first-price auctions with heterogeneous bidders: An analysis of merger effects." *American Economic Journal: Microeconomics*, 7(2): 188–214.

Maluccio, John, and Rafael Flores. 2005. "Impact evaluation of a conditional cash transfer program: The Nicaraguan Red de Protección Social." *International Food Policy Research Institute: Food Consumption and Nutrition Division Discussion Paper*, No. 184.

McFadden, Daniel. 1989. "A Method of Simulated Moments for Estimation of Discrete Response Models Without Numerical Integration." *Econometrica*, 57(5): 995.

Morten, Melanie. 2019. "Temporary migration and endogenous risk sharing in village india." *Journal of Political Economy*, 127(1): 1–46.

Nocedal, Jorge, and Stephen J. Wright. 2006. *Numerical Optimization.* Springer New York.

OECD. 2017. *Education at a Glance.* OECD.

Riccio, James, Nadine Dechausay, Cynthia Miller, Stephen Nuñez, Nandita Verma, and Edith Yang. 2013. "Conditional Cash Transfers in New York City: The Continuing Story of the Opportunity NYC-Family Rewards Demonstration." *MDRC.*

Schultz, T Paul. 2004. "School subsidies for the poor: Evaluating the Mexican PROGRESA poverty program." *Journal of Development Economics*, 74(1): 199–250.

Skira, Meghan M. 2015. "Dynamic wage and employment effects of elder parent care." *International Economic Review*, 56(1): 63–93.

Smith Jr, Anthony A. 1993. "Estimating nonlinear time-series models using simulated vector autoregressions." *Journal of Applied Econometrics*, 8(S1): S63–S84.

Ulyssea, Gabriel. 2018. "Firms, informality, and development: Theory and evidence from Brazil." *American Economic Review*, 108(8): 2015–47.

University College London, UCL Institute of Education, Centre for Longitudinal Studies. 2020. "Next Steps: Linked Education Administrative Datasets (National Pupil Database), England, 2005-2009: Secure Access. 6th Edition." UK Data Service, SN: 7104.

University College London, UCL Institute of Education, Centre for Longitudinal Studies. 2021. "Next Steps: Sweeps 1-8, 2004-2016. 15th Edition." UK Data Service, SN: 5545.

Appendix

A Details on benefits accruing to parents

The two government transfers to parents included in the model are Child Benefit (CB) and the Child Tax Credit (CTC). Like the EMA, both of these benefits were paid to parents of students in 'non-advanced education' up to age 19.[41] CB was a universal grant that was independent of parental income. Rates were higher for the eldest child in the family than for subsequent children. The relevant weekly rates for our sample period were (in current prices) £17.45 in 2006/07, £18.10 in 2007/08, £18.80 in 2008/09 for the eldest child, and £11.70 in 2006/07, £12.10 in 2007/08, £12.55 in 2008/09 for subsequent children.

CTC eligibility depended on parental income. It had two components with different eligibility thresholds: the 'family element' that was available to a majority of families, and the 'child element' that was only available to relatively poor families. The family element was frozen in nominal terms at £545 (annual) and withdrawn at a rate of 6.67 pence in the pound for families with incomes above £50,000. The child element (for non-disabled children; annual) was £1,765 in 2006/07, £1,845 in 2007/08, £2,085 in 2008/09 in current prices. It was withdrawn at a rate of 37 pence in the pound (39 pence in the pound in 2008/09) for families with incomes above a threshold of £14,155 in 2006/07, £14,495 in 2007/08 and £15,575 in 2008/09. According to Brewer (2003), around 90% of families with dependent children were eligible for some amount of CTC in the 2003/04 tax year. As eligibility thresholds for the 'family allowance' remained frozen in nominal terms, somewhat fewer families (but still a large majority) were eligible for CTC in our sample period.

B Household Bargaining

Assume that parents and their children play a sequential game, where children decide first between education ($d_{St} = 1$), seeking employment ($d_{Et} = 1$), or NEET ($d_{Nt} = 1$). If the child chooses to seek employment, she either finds or does not find work. Then the parents decide on a (non-negative) amount T_{mt} to transfer to their child, taking her activity status $\boldsymbol{s}_{mt}$ as given. Students know their parents' utility function and thus T_{mt}, and will take this into account when

[41] For those reaching their 19th birthday while enrolled on a course, CB and CTC payments were continued until the end of the course, i.e. until the end of the academic year in most cases.

choosing their action.[42]

Assume parents' utility is given by

$$U_{mt}^{P} = \log C_{mt}^{P} + \chi \left[\log C_{mt} + \delta^{P} E(V_{t+1}(\Omega_{t+1}) | \boldsymbol{s}_{mt} = 1) \right]. \tag{B.1}$$

where C_{mt}^{P} is parental consumption if the child's choice leads to $s_{mt} = 1$. Parents are assumed to care about their child's well-being: The term in square brackets represents the child's utility. Note that because children move first, it is irrelevant for the parents' problem whether the parent has a different view of the appropriate discount rate from the child, i.e. $\delta^{P} \neq \delta$. The parents' budget constraint is

$$C_{mt}^{P} = Y^{P} + B_{mt}^{P} - T_{mt}. \tag{B.2}$$

The solution of the parents' problem of choosing T_{mt} to maximize their utility function given by (B.1) subject to (B.2) and the child's budget constraint is (36). This implies that as long as the child receives a positive transfer, income from all sources is counted the same, irrespective to whom in accrues. In that case, the child will always consume a constant share of total family income governed by χ.

One way to break this equivalence result is to assume that while parents correctly perceive their child's (flow) utility function, they take their child's budget constraint to be

$$C_{mt} = \mu Y_{mt} + T_{mt} \tag{B.3}$$

which includes the actual budget constraint (31) as a special case in which $\mu = 1$. One rationale for $\mu < 1$ might be that children can systematically hide income from their parents. Another is that parents believe children will invariably waste some of the income they receive on frivolous expenses, whereas in-kind transfers from parents (e.g. housing or meals) are non-wastable. With $\mu < 1$, a transfer paid to the child like the EMA would have a larger effect than a transfer to the parents.[43]

However, we found no evidence for $\mu < 1$ in our context. In fact, when we estimated μ along with other model parameters, the best fit was obtained with $\mu > 1$, suggesting that

[42] This modelling choice is motivated by the fact that students can to some extent commit to a course of action (e.g. incurring a job search cost or failing to register for the new school year), whereas parents cannot.

[43] Who receives a transfer has been shown to have a large effect in a developing country context. For instance, Attanasio, Meghir and Santiago (2012) found that allowing for a differential utility payoff from a Mexican CCT grant roughly doubled the estimated effect of the programme.

if anything, parents may withdraw transfers at a higher rate. As the parameter was not well-identified, we have retained the assumption that parents correctly perceive their child's budget constraint, corresponding to $\mu = 1$.

C Auxiliary Model Parameters

Here we present analytical forms for the parameters of both the unsmoothed and the smoothed version of our auxiliary model. In contrast to the main text, individual level variables are indexed by i. In contrast to section 4.4, we do not explicitly indicate dependence on Ω_{it}^* or $\boldsymbol{\rho}$ to avoid notational clutter.

The unsmoothed status shares by EMA eligibility groups are given by

$$\mathcal{S}_{kt}^g = \frac{\sum_{i=1}^N W_i^{samp} I(G_i = g) s_{kit}}{\sum_{i=1}^N W_i^{samp} I(G_i = g)} \tag{C.1}$$

where g indexes the different EMA eligibility groups (£30, £20, £10, ineligible), and W_i^{samp} is the sampling weight (including age 16 attrition weights) for individual i taken from *Next Steps*. Smoothed status shares are

$$\tilde{\mathcal{S}}_{kt}^g = \frac{\sum_{i=1}^N W_i^{samp} I(G_i = g) \sum_{\mathfrak{s}_{t-1}^* \in S_{t-1}} W_{kit}^s(\mathfrak{s}_{i,t-1}^*)}{\sum_{i=1}^N W_i^{samp} I(G_i = g)}. \tag{C.2}$$

The unsmoothed unemployment shares by EMA eligibility group are given by

$$u_t^g = \frac{\sum_{i=1}^N W_i^{samp} I(G_i = g)(1 - j_{it}) c_{Eit}}{\sum_{i=1}^N W_i^{samp} I(G_i = g) s_{Nit}} \tag{C.3}$$

Smoothed unemployment shares are

$$\tilde{u}_t^g = \frac{\sum_{i=1}^N W_i^{samp} I(G_i = g) \sum_{\mathfrak{s}_{t-1}^* \in S_{t-1}} \left(1 - W^j(\mathfrak{s}_{i,t-1}^*)\right) W_{Eit}^c(\mathfrak{s}_{i,t-1}^*)}{\sum_{i=1}^N W_i^{samp} I(G_i = g) \sum_{\mathfrak{s}_{t-1}^* \in S_{t-1}} W_{Nit}^s(\mathfrak{s}_{i,t-1}^*)}. \tag{C.4}$$

Parent transfer shares at age 26 are given by

$$\mathcal{T}^g = \frac{\sum_{i=1}^N W_i^{samp} I(G_i = g) I(T_{iT} > 0)}{\sum_{i=1}^N W_i^{samp} I(G_i = g)} \tag{C.5}$$

Smoothed age 26 transfer shares are given by

$$\tilde{\mathcal{T}}^g = \frac{\sum_{i=1}^N W_i^{samp} I(G_i = g) \sum_{\mathfrak{s}^*_{T-1} \in S_{T-1}} \sum_{k \in \{S,E,N\}} W^s_{kiT}(\mathfrak{s}^*_{i,T-1}) I(T_{kiT} > 0)}{\sum_{i=1}^N W_i^{samp} I(G_i = g)} \tag{C.6}$$

Average wages by EMA eligibility group are

$$\overline{w}_t^g = \frac{\sum_{i=1}^N W_i^{samp} I(G_i = g) s_{Eit} \log w_{it}}{\sum_{i=1}^N W_i^{samp} I(G_i = g) s_{Eit}}. \tag{C.7}$$

where the sample is divided into just two groups (eligible and ineligible). The smooth analogue is

$$\tilde{w}_t^g = \frac{\sum_{i=1}^N W_i^{samp} I(G_i = g) \sum_{\mathfrak{s}^*_{t-1} \in S_{t-1}} W^s_{Eit}(\mathfrak{s}^*_{i,t-1}) \log w_{it}(\mathfrak{s}^*_{i,t-1})}{\sum_{i=1}^N W_i^{samp} I(G_i = g) \sum_{\mathfrak{s}^*_{t-1} \in S_{t-1}} W^s_{Eit}(\mathfrak{s}^*_{i,t-1})}. \tag{C.8}$$

The (j,k)-entry of the unsmoothed transition matrix of transitions to period t for EMA eligibility group g is

$$M^g_{tjk} = \frac{\sum_{i=1}^N W_i^{samp} I(G_i = g) s_{ji,t-1} s_{kit}}{\sum_{i=1}^N W_i^{samp} I(G_i = g) s_{ji,t-1}}. \tag{C.9}$$

The smoothed analogue is

$$\tilde{M}^g_{tjk} = \frac{\sum_{i=1}^N W_i^{samp} I(G_i = g) \sum_{\mathfrak{s}^*_{t-1} \in S_{t-1}} W^s_{kit}(\mathfrak{s}^*_{i,t-1}) \tilde{s}_{ji,t-1}(\mathfrak{s}^*_{i,t-1})}{\sum_{i=1}^N W_i^{samp} I(G_i = g) \sum_{\mathfrak{s}^*_{t-1} \in S_{t-1}} W_{it}(\mathfrak{s}^*_{i,t-1}) \tilde{s}_{ji,t-1}(\mathfrak{s}^*_{i,t-1})}. \tag{C.10}$$

where

$$\tilde{\boldsymbol{s}}_{it}(\mathfrak{s}^*_{it}) = \begin{cases} \left(1 - \frac{W^s_{Nit}(\mathfrak{s}^*_{i,t-1})}{W^{s*}_{Nit}(\mathfrak{s}^*_{i,t-1})}, 0, \frac{W^s_{Nit}(\mathfrak{s}^*_{i,t-1})}{W^{s*}_{Nit}(\mathfrak{s}^*_{i,t-1})}\right) & \text{if } s^*_{Nit} = 1 \\ \boldsymbol{s}^*_{it} & \text{otherwise.} \end{cases} \tag{C.11}$$

The reason $\boldsymbol{s}^*_{it}$ needs to be adjusted in this way is that $s^*_{Nit} = 1$ will also hold for cases where an individual chose education but did not pass the grade, which will be recorded as 'in education' in the data. The adjustment above implies that $\tilde{\boldsymbol{s}}_{it}(\mathfrak{s}^*_{it})$ need not be a vector of indicator variables.

For the wage equation, the unsmoothed OLS parameter estimates are

$$\boldsymbol{\beta}^* = (X^{W\prime} X^W)^{-1} X^{W\prime} \boldsymbol{y}^W \tag{C.12}$$

where each row $\boldsymbol{x}^W$ of X^W is

$$\boldsymbol{x}^W = (1, \log Y^p_i, F_i, A_{11,i}, A_{16,i}, X_{i,t-1}, S_{i,t-1}) \tag{C.13}$$

and each element of $\boldsymbol{y}^W$ is an individual's wage w_{it}. Rows are included in X^W if and only if $s_{Eit} = 1$ and $t \in \{1, 2, 3, 4, 10\}$, as wages are only observed in these periods. The smooth analogue is the weighted least squares regression

$$\tilde{\boldsymbol{\beta}}^* = (\tilde{X}^{W\prime}\tilde{X}^W)^{-1}\tilde{X}^{W\prime}\tilde{\boldsymbol{y}}^W \tag{C.14}$$

where each row of $\tilde{X}^W$ is

$$\tilde{\boldsymbol{x}}^W(\mathfrak{s}^*_{i,t-1}) = (1, \log Y_i^P, F_i, A_{11,i}, A_{16,i}, X_{i,t-1}(\mathfrak{s}^*_{i,t-1}), S_{i,t-1}(\mathfrak{s}^*_{i,t-1}))\sqrt{W_i^{samp} W^s_{Eit}(\mathfrak{s}^*_{i,t-1})} \tag{C.15}$$

and

$$\tilde{y}^W(\mathfrak{s}^*_{i,t-1}) = w_{it}(\mathfrak{s}^*_{i,t-1})\sqrt{W_i^{samp} W^s_{Eit}(\mathfrak{s}^*_{i,t-1})}. \tag{C.16}$$

$$X_{i,t-1}(\mathfrak{s}^*_{i,t-1}) = \sum_{\tau=1}^{t-1} \tilde{s}_{Ei\tau}(\mathfrak{s}^*_{i\tau}) = \sum_{\tau=1}^{t-1} s^*_{Ei\tau} \tag{C.17}$$

and

$$S_{i,t-1}(\mathfrak{s}^*_{i,t-1}) = \sum_{\tau=1}^{t-1} \tilde{s}_{Si\tau}(\mathfrak{s}^*_{i\tau}) \tag{C.18}$$

with $\tilde{\boldsymbol{s}}_{it}(\mathfrak{s}^*_{it})$ defined as in (C.11). In principle, values corresponding to all *possible* values of $\mathfrak{s}^*_{i,t-1}$ are included in the regression; in practice, only those with positive weight need to be included. Again only values with time subscripts of $t \in \{1, 2, 3, 4, 10\}$ are included to match the data.

The linear probability model of job finding is estimated analogously. OLS estimates are

$$\boldsymbol{\lambda}^* = (X^{J\prime}X^J)^{-1}X^{J\prime}y^J \tag{C.19}$$

where each row $\boldsymbol{x}^J_l$ of X^J is

$$\boldsymbol{x}^J = (1, \log Y_i^P, F_i, A_{11,i}, A_{16,i}, S_{i,t-1}, X_{i,t-1}, s_{1i,t-1}, s_{2i,t-1}) \tag{C.20}$$

and $y^J = j_{it}$. Rows are included if and only if $c_{Eit} = 1$. The smooth analogue is the weighted least squares regression

$$\tilde{\boldsymbol{\lambda}}^* = (\tilde{X}^{J\prime}\tilde{X}^J)^{-1}\tilde{X}^{J\prime}\tilde{\boldsymbol{y}}^J \tag{C.21}$$

where each row of $\tilde{X}^J$ is

$$\tilde{\boldsymbol{x}}^J(\mathfrak{s}^*_{i,t-1}) = \\ (1, \log Y_i^P, F_i, A_{11,i}, A_{16,i}, S_{i,t-1}(\mathfrak{s}^*_{i,t-1}), X_{i,t-1}(\mathfrak{s}^*_{i,t-1}), \tilde{s}_{1i,t-1}(\mathfrak{s}^*_{i,t-1}), \tilde{s}_{2i,t-1}(\mathfrak{s}^*_{i,t-1}))\sqrt{W_i^{samp} W_{Eit}^c(\mathfrak{s}^*_{i,t-1})} \tag{C.22}$$

and

$$\tilde{y}^J(\mathfrak{s}^*_{i,t-1}) = j_{it}(\mathfrak{s}^*_{i,t-1})\sqrt{W_i^{samp} W_{Eit}^c(\mathfrak{s}^*_{i,t-1})}. \tag{C.23}$$

Again, values corresponding to all *possible* values of $\mathfrak{s}^*_{i,t-1}$ are included in the regression in principle, and only those with positive weight in practice.

D Further Parameter Estimates

Table 6 lists the remaining model parameters not given in Table 3 above. The top panel lists the parameter that governs resource sharing between parents and children, χ; the estimated tuition cost, τ; the re-entry cost to education, γ^r; and the job search cost, γ^{js}. χ was estimated in logs to ensure positivity.

Table 6: Other model parameters

Parameter	Estimate	Asymptotic Standard Error
$\log \chi$	-0.976	0.145
τ	-16670.457	3468.059
γ^r	4.274	0.467
γ^{js}	0.080	0.038
$\sqrt{\kappa_{11}}$	0.982	0.047
κ_{21}	0.678	0.078
$\sqrt{\kappa_{22}}$	1.127	0.039
κ_{31}	-0.516	0.101
κ_{32}	-0.431	0.069
$\sqrt{\kappa_{33}}$	0.687	0.058

Notes. χ was estimated in logs and κ_{11}, κ_{22} and κ_{33} were estimated as square roots to ensure positivity.

The bottom panel relates to the covariance matrix of the shock vector $\boldsymbol{e}_t = (\epsilon_t, \omega_t, \eta_t)'$, Σ. We parameterize Σ, the covariance matrix of the shock vector $\boldsymbol{e}_t = (\epsilon_t, \omega_t, \eta_t)'$, using the

Cholesky decomposition. In particular, we estimate the parameters of

$$K = \begin{pmatrix} \kappa_{11} & & \\ \kappa_{21} & \kappa_{22} & \\ \kappa_{31} & \kappa_{32} & \kappa_{33} \end{pmatrix} \tag{D.1}$$

where $KK^T = \Sigma$. Table 6 gives the estimates of the κs. The entries along the main diagonal κ_{11}, κ_{22} and κ_{33} were estimated as square roots to ensure positivity.

The estimated parameter value for $\log \chi$ implies that if there is a positive transfer, parents allocate slightly more than a quarter of household income to their children. The tuition fee parameter τ is significantly negative. This result might be due to the generosity of the UK grants system for university students at the time, or may reflect the particular value of advanced education that is not captured by the model.

As might be expected, the cost of re-entry into education or training γ^r and the job search cost γ^{js} are positive. The re-entry cost to education in particular implies a very high cost of re-entry. The estimates of the κs imply a moderate variance of the shocks; all have roughly similar variance.

E Regression parameters used to assign age 11 and age 16 attainment

Table 7: Age 16 attainment: at least five A*-C at GCSE (Logit model)

Regressor	**Estimate**	**Standard Error**
Parent Income Band 2: £10,400 to £15,599	0.496	0.097
Parent Income Band 3: £15,600 to £20,799	0.630	0.108
Parent Income Band 4: £20,800 to £25,999	0.903	0.107
Parent Income Band 5: £26,000 to £31,199	1.105	0.105
Parent Income Band 6: £31,200 to £41,799	1.395	0.094
Parent Income Band 7: £41,800 to £51,999	1.683	0.109
Parent Income Band 8: £52,000 or more	2.079	0.120
Eldest child	0.460	0.057
Female	0.462	0.058
At least one parent with degree	1.175	0.092

Notes. Logistic regression of an indicator for obtaining at least five A*-C at GCSE on parental income band and a number of other background characteristics ($N = 8,706$). The regression was weighted to take account of stratified sampling and attrition. Standard errors take into account the complex sampling scheme of the *Next Steps* survey. The intercept term was suppressed for statistical disclosure reasons.

Table 8: Age 11 attainment: KS2 score (linear model)

Regressor	**Estimate**	**Standard Error**
Parent Income Band 2: £10,400 to £15,599	0.540	0.207
Parent Income Band 3: £15,600 to £20,799	0.856	0.198
Parent Income Band 4: £20,800 to £25,999	0.713	0.207
Parent Income Band 5: £26,000 to £31,199	1.146	0.200
Parent Income Band 6: £31,200 to £41,799	1.237	0.176
Parent Income Band 7: £41,800 to £51,999	1.230	0.193
Parent Income Band 8: £52,000 or more	1.733	0.181
Eldest child	0.309	0.074
Female	-0.136	0.079
At least one parent with degree	0.785	0.102
At least five A*-C at GCSE	4.737	0.122

Notes. Linear regression of an KS2 score on parental income band and a number of other background characteristics, including an indicator for obtaining at least 5 A*-C marks at GCSE ($N = 8,433$). The regression was weighted to take account of stratified sampling and attrition. Standard errors take into account the complex sampling scheme of the *Next Steps* survey. The intercept term was suppressed for statistical disclosure reasons. The root mean squared error of the predicted values was $\sigma = 2.9982$.

III. Revisiting the solution of dynamic discrete choice models: time to bring back Keane and Wolpin (1994)?*

JACK BRITTON
IFS

BEN WALTMANN
Oxford and IFS

Abstract

The 'curse of dimensionality' is a common problem in the estimation of dynamic models: as models get more complex, the computational cost of solving these models rises exponentially. Keane and Wolpin (1994) proposed a method for addressing this problem in finite-horizon dynamic discrete choice models by evaluating only a subset of state space points by Monte Carlo integration and interpolating the value of the remainder. This method was widely used in the late 1990s and 2000s but has rarely been used since, as it was found to be unreliable in some settings. In this paper, we develop an improved version of their method that relies on three amendments: systematic sampling, data-guided selection of state space points for Monte Carlo integration, and dispensing with polynomial interpolation when a multicollinearity problem is detected. With these improvements, the Keane and Wolpin (1994) method achieves excellent approximation performance even in a model with a large state space and substantial *ex ante* heterogeneity.

*We are grateful to Rob Sauer for many helpful discussions on this topic, and to James Duffy, Zoe Fannon, Mike Keane, Hamish Low, and seminar participants at Oxford for useful comments and suggestions. The use of the University of Oxford Advanced Research Computing (ARC) facility is gratefully acknowledged (DOI code: http://dx.doi.org/10.5281/zenodo.22558).

1 Introduction

Many economic models are *dynamic* in the sense that current choices affect not only current but also expected future returns. A common problem in the estimation of these models is the 'curse of dimensionality': as models get more complex, the computational cost of solving them rises exponentially with the dimensionality of the state space. Keane and Wolpin (1994) (hereafter KW94) proposed a very general method for addressing this problem for the subclass of dynamic models where agents repeatedly choose from a discrete set of options and the time horizon is finite. Models with this structure are common in labour economics, where they have been used to study career choice and labour supply.

The KW94 method works backwards from the final period. A subset of state space points is evaluated by Monte Carlo integration each period. The value of the remaining points is imputed using a polynomial interpolation method. This saves computation time relative to a full solution, where all state space points are evaluated using Monte Carlo integration. The method was widely used in empirical work in the late 1990s and 2000s.[1] However, it has rarely been used since, as it was found to be unreliable in some settings.[2]

In this paper, we develop an improved version of the KW94 method. Using two versions of the KW94 model as test cases, we show that the improved KW94 method performs very well, both in Monte Carlo simulations at the true parameter values and when the model parameters are estimated from simulated data. Holding the number of integral simulation draws constant, simulated choices of individuals when the improved KW94 approximation is used are nearly as close to 'true' simulated choices as when the full solution is used. When the model is estimated from simulated data, this translates into accurate parameter estimates and precise predictions in a simple policy experiment, again in line with the full solution. These results obtain both in the canonical KW94 model and in a variation of this model with a larger state space and substantial *ex ante* heterogeneity, where the traditional version of the KW94 method is unreliable.

At the same number of integral simulation draws, our improved KW94 approximation

[1] Important papers in labour economics using the method include Keane and Wolpin (1997), Keane and Wolpin (2001), Imai and Keane (2004), Blau and Gilleskie (2006), Lee and Wolpin (2006), Van der Klaauw and Wolpin (2008), Lee and Wolpin (2010) and Keane and Wolpin (2010). It has also been applied in other fields, including marketing science (Erdem and Keane, 1996), health economics (Crawford and Shum, 2005) and development economics (Todd and Wolpin, 2006).

[2] Nevertheless, there is enduring interest in the KW94 approach: in a recent contribution, Eisenhauer (2019) reproduces the results in KW94 and provides an open-source Python package for implementing the method.

is much faster to compute than a full solution; in the two versions of the KW94 model we look at in this paper, the difference is around an order of magnitude. When we instead hold computation time constant, the improved KW94 method is nearly always more accurate than the full solution except at very high computation times. The reason is that our improved KW94 method focuses computational resources on the most important state space points, which is generally preferable to spreading resources equally. This substantially improves the trade-off between model complexity, accuracy, and computation time for applied researchers.

Our improved version of the KW94 method is based on three amendments to the traditional method. First, we use a systematic sampling method instead of simple random sampling for Monte Carlo integration. Systematic sampling reduces the error in integral simulation for the random subset of state space points that are evaluated by Monte Carlo integration. It also indirectly leads to more accurate interpolation, as the parameters of the interpolation polynomial can be more precisely estimated.

Second, taking a data-guided approach to selecting state space points for evaluation by Monte Carlo integration is preferable to the approach recommended in KW94 of selecting them at random. In particular, drawing state space points for Monte Carlo evaluation from the subset of points that agents reach in the data — with the probability of any particular state space point being evaluated proportional to the number of agents reaching that point — can allow for a degree of accuracy comparable to a full model solution even with a very small number of state space points evaluated. Especially when combined, these two improvements can lead to large reductions in required computation time for a desired level of accuracy.

Third, even with these two amendments, the KW94 method can still be unreliable for models with substantial *ex ante* heterogeneity unless a large number of state space points is evaluated by numerical integration. The reason is that unlike in the canonical KW94 model, the estimation of the interpolation polynomial in these models frequently suffers from a multicollinearity problem. We show that a simple and effective solution to this problem is to dispense with the interpolation step if a substantial multicollinearity problem is detected. Instead, it is better to rely solely on an approximation provided by Jensen's Inequality in these cases.

We conclude that the KW94 method, thus improved, is still a useful tool for applied researchers seeking to estimate finite-horizon discrete choice dynamic programming models with large state spaces. It is especially valuable in cases where the model structure precludes the use of the Conditional Choice Probability (CCP) approach of Hotz and Miller (1993) and

Arcidiacono and Miller (2011). This is commonly the case in labour economics, as the CCP approach generally requires error terms to be additive. Models with non-additive error terms include KW94 and Keane and Wolpin (1997).

The rest of the paper is structured as follows. Section 2 briefly lays out the canonical model of KW94. Section 3 outlines our improvements to the KW94 method and provides Monte Carlo evidence of the gains in approximation performance associated with them in the KW94 model. Section 4 shows that these differences in approximation performance translate into differences in the accuracy of parameter estimates and counterfactual predictions. Section 5 introduces *ex ante* heterogeneity into the KW94 model and investigates the performance of the different versions of the KW94 method relative to the full solution in that setting. Section 6 concludes.

2 A canonical model

The model of KW94 is a model of occupational choice. In each of $T = 40$ discrete periods of time, agents choose between $K = 4$ different options: blue-collar work, white-collar work, education, and home production. The per-period utility functions are

$$u_{1t} = w_{1t} = \exp(\alpha_{10} + \alpha_{11}s_t + \alpha_{12}x_{1t} - \alpha_{13}x_{1t}^2 + \alpha_{14}x_{2t} - \alpha_{15}x_{2t}^2 + \epsilon_{1t}) \tag{1}$$

$$u_{2t} = w_{2t} = \exp(\alpha_{20} + \alpha_{21}s_t + \alpha_{22}x_{2t} - \alpha_{23}x_{2t}^2 + \alpha_{24}x_{1t} - \alpha_{25}x_{1t}^2 + \epsilon_{2t}) \tag{2}$$

$$u_{3t} = \beta_0 - \beta_1 I(s_t \geq 13) - \beta_2(1 - d_{3,t-1}) + \epsilon_{3t} \tag{3}$$

$$u_{4t} = \gamma_0 + \epsilon_{4t} \tag{4}$$

w_{1t} and w_{2t} are the agent's (latent) wages in occupation one and two. s_t is the number of periods of schooling accumulated by the beginning of period t. x_{1t} and x_{2t} are an individual's total periods of work experience at the beginning of period t in occupation one and two, respectively. $\boldsymbol{d}_t$ is a vector of indicator variables, where for each element $d_{kt} = 1$ if option k is chosen in period t and $d_{kt} = 0$ otherwise. Hence $d_{3,t-1}$ is an indicator variable of whether schooling was chosen in the previous period. $\boldsymbol{\epsilon}_t \sim N(0, \Sigma)$ is a vector of serially uncorrelated shocks, where Σ is parameterized by the parameter vector $\boldsymbol{a}$. $\boldsymbol{\epsilon}_t$ is known to the agent at the beginning of period t but not before. $\boldsymbol{\theta} = \{\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{a}\}$ is the full parameter vector of the model.

Individuals have perfect knowledge of the true model and rational expectations about the future. Let $\boldsymbol{S}_t = \{s_t, x_{1t}, x_{2t}, d_{3,t-1}, \boldsymbol{\epsilon}_t\}$ be the vector of state variables, and denote the vector of pre-determined state variable by $\overline{\boldsymbol{S}}_t = \{s_t, x_{1t}, x_{2t}, d_{3,t-1}\}$. The vector of pre-determined state variables evolves deterministically given the previous period's choices: $x_{1,t+1} = x_{1,t} + d_{1,t}$, $x_{2,t+1} = x_{2,t} + d_{2,t}$, $s_{t+1} = s_t + d_{3,t}$, and $d_{3,t-1}$ is simply the previous period's choice. Initial conditions are $x_{11} = x_{21} = 0$ and $s_1 = 10$. All agents are *ex ante* homogeneous; different choices arise only due to different draws of $\boldsymbol{\epsilon}_t$.[3]

Then the agent's decision problem in period t can be written recursively as

$$\underset{k \leq K}{\operatorname{argmax}}\, V_{kt}(\boldsymbol{S}_t) \tag{5}$$

where

$$V_{kt}(\boldsymbol{S}_t) = \begin{cases} u_{kt}(\boldsymbol{S}_t) + \delta \, \mathrm{E}\left[\max_{j \leq K} V_{j,t+1}(\boldsymbol{S}_{t+1}) | \overline{\boldsymbol{S}}_t, d_{kt} = 1\right] & \text{if } t < T \\ u_{kt}(\boldsymbol{S}_t) & \text{if } t = T \end{cases} \tag{6}$$

is the *alternative-specific value function*. E denotes the mathematical expectations operator. The discount factor δ is fixed at $\delta = 0.95$.

3 Numerical solution of the canonical model

The agent's problem in this model can be solved by backward recursion using (5) and (6). The key challenge is the evaluation of the conditional expectation in (6), which is often called 'EMAX', as it is an expectation of a maximum. This expectation is taken over the joint distribution of $\boldsymbol{\epsilon}_t$ and takes the form:

$$\underbrace{\mathrm{E}\left[\max_{k \leq K} V_{kt}(\boldsymbol{S}_t)\right]}_{\text{EMAX}} = \int \max_{k \leq K} V_{kt}(\overline{\boldsymbol{S}}_t, \boldsymbol{\epsilon}_t) d\Phi(\boldsymbol{\epsilon}_t) \tag{7}$$

where Φ is the (multivariate normal) cdf of $\boldsymbol{\epsilon}_t$ and $\boldsymbol{S}_t$ is written as $(\overline{\boldsymbol{S}}_t, \boldsymbol{\epsilon}_t)$ to emphasize the dependence on $\boldsymbol{\epsilon}_t$.

As there is in general no analytical solution for this integral, it usually has to be evaluated using numerical integration methods at considerable computational cost. It is commonly evaluated using crude Monte Carlo integration. In particular, each expectation is replaced by the

[3] A version of this model that features *ex ante* heterogeneity in the form of different per-period utility functions for different types of agents is introduced in Section 5.

consistent and unbiased estimator:

$$\underbrace{\hat{\mathrm{E}}\left[\max_{k\leq K} V_{kt}(\boldsymbol{S}_t)\right]}_{\widehat{EMAX}} = \frac{1}{D}\sum_{d=1}^{D} \max_{k\leq K} V_{kt}(\overline{\boldsymbol{S}}_t, \boldsymbol{\epsilon}_t^d) \tag{8}$$

where the random vectors $\boldsymbol{\epsilon}_t^d$ are drawn from their assumed distribution in the model and D is the number of simulation draws. This procedure is computationally demanding, as a large number of draws have to be evaluated to ensure reasonable accuracy of the numerical integration procedure, and 163,409 expectations need to be evaluated in total.[4] While modern computers can perform this operation in seconds, computational speed is still an important concern, as the estimation of model parameters typically requires the model to be solved many thousands of times at different parameter values.

KW94 offer a method for speeding up the solution of models of this type. They propose evaluating only a random subset of expectations in each period by Monte Carlo integration. All other expectations are approximated by

$$\underbrace{\tilde{\mathrm{E}}[\max_{k\leq K} V_{kt}(\boldsymbol{S}_t)]}_{\widetilde{\mathrm{EMAX}}} = \underbrace{\max_{k\leq K} \mathrm{E}\left[V_{kt}(\boldsymbol{S}_t)\right]}_{\mathrm{MAXE}} + g\left(\tilde{V}_{1t}, \ldots, \tilde{V}_{Kt}\right) \tag{9}$$

where 'MAXE' is obtained by exchanging the order of the expectation and max operators, g is an interpolation function, and

$$\tilde{V}_{jt}(\boldsymbol{S}_t) = \underbrace{\max_{k\leq K} \mathrm{E}\left[V_{kt}(\boldsymbol{S}_t)\right]}_{\mathrm{MAXE}} - \mathrm{E}\left[V_{jt}(\boldsymbol{S}_t)\right] \text{ for } j = 1, \ldots, K. \tag{10}$$

As recommended in KW94, it is natural to impose $g(.) \geq 0$, since by Jensen's Inequality, $\max_{k\leq K} \mathrm{E}\left[V_{kt}(\boldsymbol{S}_t)\right]$ ('MAXE') is a lower bound for $\mathrm{E}\left[\max_{k\leq K} V_{kt}(\boldsymbol{S}_t)\right]$ ('EMAX').[5]

In KW94's preferred specification, the interpolation function g is parameterized as

$$g\left(\tilde{V}_{1t}(\boldsymbol{S}_t), \ldots, \tilde{V}_{Kt}(\boldsymbol{S}_t)\right) = \pi_0 + \sum_{k=1}^{K} \pi_k \tilde{V}_{kt}(\boldsymbol{S}_t) + \sum_{k=1}^{K} \pi_{K+k} \sqrt{\tilde{V}_{kt}(\boldsymbol{S}_t)}. \tag{11}$$

[4] This number reflects the total number of possible combinations of state variables, added up over 39 periods ($t = 2$ to $t = 40$), and taking into account that the maximum permissible number of years of schooling is $s_{max} = 20$, and that it is not possible to have $d_{3,t-1} = 1$ and $s_t = s_0$ or $d_{3,t-1} = 0$ and $s_t = s_0 + t - 1$.

[5] Intuitively, MAXE represents the value of a 'Plan A', whereas EMAX takes into account the value of 'Plans B to D'. As shown in Appendix B, tighter analytical bounds for EMAX are available, but imposing them leads to minor gains at best.

The parameters of this interpolation function are estimated from the subset of expectations evaluated by Monte Carlo integration in each period of the recursion. They are obtained by OLS regression, with

$$y(\boldsymbol{S}_t) = \underbrace{\hat{\mathrm{E}}[\max_{k \leq K} V_{kt}(\boldsymbol{S}_t)]}_{\widehat{EMAX}} - \underbrace{\max_{k \leq K} \mathrm{E}\left[V_{kt}(\boldsymbol{S}_t)\right]}_{\mathrm{MAXE}}$$

as the dependent variable.[6]

As shown in KW94 and confirmed in section 3.4, this method generally provides a good approximation to the full model solution at much lower computational cost. However, the accuracy of the approximation deteriorates substantially as the number of state space points evaluated by Monte Carlo integration is reduced. The method can also be unreliable in some settings; an important case is the estimation of models that feature substantial *ex ante* heterogeneity.

We propose three amendments to the method that address these issues. The first is to use a systematic sampling method for Monte Carlo integration instead of simple random sampling. The second is to select expectations to be evaluated by Monte Carlo integration not at random but based on the part of the state space that agents actually reach in the data. The third detects and avoids multicollinearity problems in the interpolation regression. Together, these changes dramatically improve the approximation performance of the KW94 method, as we show in the remainder of this paper.[7]

3.1 Improvement 1: systematic sampling

The use of systematic sampling can lead to large improvements in the approximation performance of the KW94 method.[8] Systematic sampling guarantees better coverage of the integrals evaluated using Monte Carlo simulation. As a result, the approximation error is generally $O(D^{-1})$ instead of $O_p(D^{-1/2})$ for simple random sampling, where D is the number of simulation draws (Geweke, 1996). This increases the efficiency not only of the Monte Carlo simulator

[6] More specifically, $\boldsymbol{\pi} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$ where $\boldsymbol{X}$ is the data matrix corresponding to (11), with the number of rows of $\boldsymbol{X}$ equal to the number of state space points evaluated by Monte Carlo integration. This presupposes that $\boldsymbol{X}'\boldsymbol{X}$ is invertible, which is not guaranteed. The case when $\boldsymbol{X}'\boldsymbol{X}$ is (near-)singular is especially relevant in models with *ex ante* heterogeneity, and is discussed in section 3.3. In the canonical model, (near-)singular moment matrices $\boldsymbol{X}'\boldsymbol{X}$ almost never occur outside periods $t = 2$ and $t = 3$ at the true parameters. In those early periods, the number of possible state space points is low enough for all to be evaluated by Monte Carlo integration, obviating the need for interpolation.

[7] Appendix C shows that our improved KW94 method is also preferable to various other methods for approximating EMAX.

[8] As shown in Appendix D.2, it can also substantially improve the performance of the full solution method.

$\widehat{EMAX}$ given by (8), but also of the interpolation estimator $\widetilde{\text{EMAX}}$ given by (9), as the dependent variable of the interpolation regression will be subject to less error.

The systematic sampling algorithm used in the rest of the paper is:

1. Draw u_{kt} from the standard uniform distribution for each $k \leq K$ and $t \in [2, T]$.

2. Calculate

$$\left(\eta_{kt}^d\right)_{d=1}^D = \Phi^{-1}\left(\frac{d - u_{kt}}{D}\right)$$

   for each $k \leq K$ and $t \in [2, T]$.

3. Randomly permute the elements of each $\left(\eta_{kt}^d\right)_{d=1}^D$ to obtain $\left(\tilde{\eta}_{kt}^d\right)_{d=1}^D$

4. Obtain $\boldsymbol{\epsilon}_t^d = \boldsymbol{A}\tilde{\boldsymbol{\eta}}_t^d$, where $\boldsymbol{A}$ is the lower triangular Cholesky decomposition of $\boldsymbol{\Sigma}$.

Other variance reduction techniques are likely to have similar positive effects on approximation performance. In particular, as we show in Appendix D, using Halton (1964) draws instead of our systematic sampling scheme even leads to slightly better performance. The advantages of our simple scheme are, first, that it is guaranteed to be unbiased, second, that draws are guaranteed to be uncorrelated across both options and periods, and third, that the algorithm is simple to implement for any number of options and periods. The comparison with Halton draws suggests that the cost of these advantages in terms of approximation performance is small.[9]

3.2 Improvement 2: data-driven selection of points for Monte Carlo integration

Our second amendment to the KW94 method is the data-driven selection of state space points to be evaluated using Monte Carlo integration. The method for selecting nodes that we have found to work best is to select nodes for numerical integration that agents actually reach in the data, with more commonly reached nodes selected with a higher probability. Intuitively, the value of nodes that are commonly reached in the data is likely to be more important than that of nodes that are never actually chosen, some of which may not be "real options" given the payoff structure of the model.

A number of different selection schemes will achieve this basic objective. The particular algorithm used in the rest of the paper is:

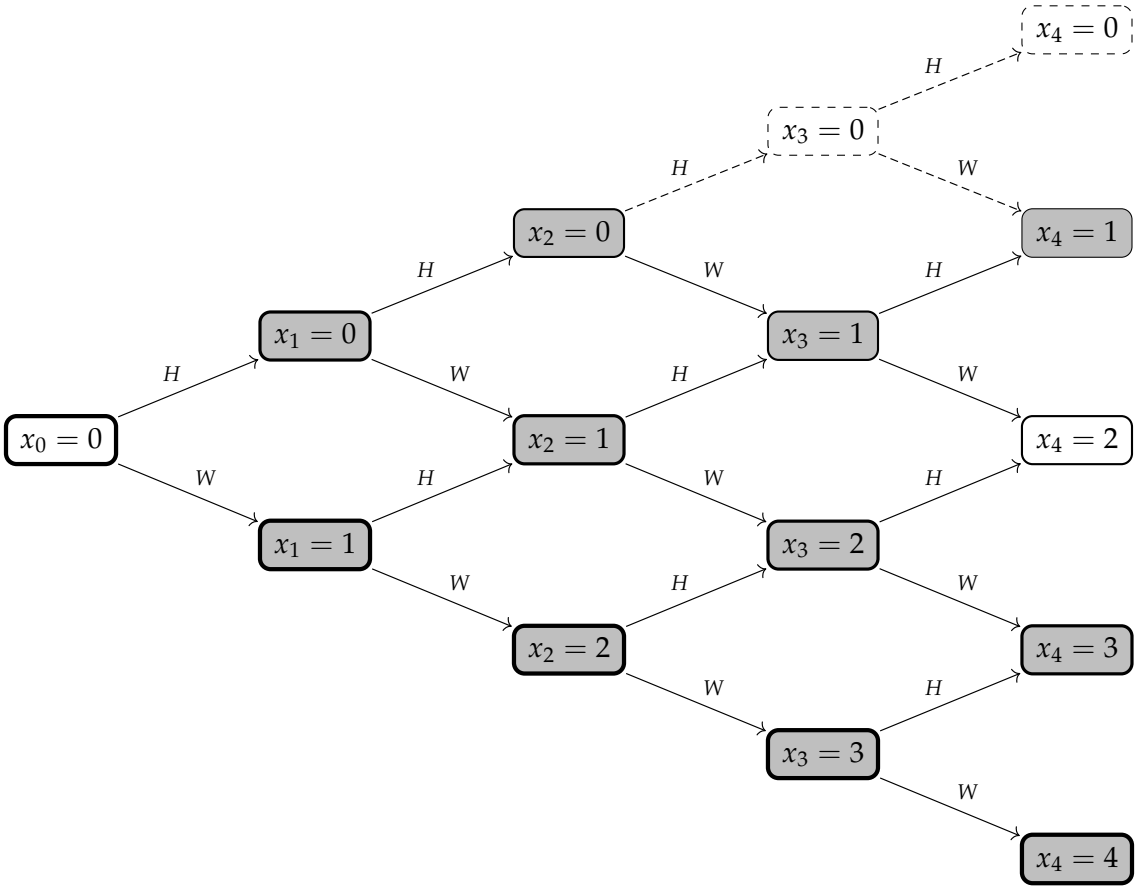
[9] For an intuitive discussion of how systematic sampling and other variance reduction techniques improve on simple random sampling, see Section 9.3 in Train (2009).

1. Permute the data so that individuals appear in random order.

2. In each period, evaluate the nodes reached in the data, working in the order that individuals appear in the data. Continue until all nodes reached in the data have been evaluated, or the maximum number of nodes per period to be evaluated by Monte Carlo integration has been reached.

3. In the former case, draw additional nodes to be evaluated by Monte Carlo integration at random from the remaining set.[10]

Figure 1: State space map for a model with $K = 2$ and $T = 5$



Notes. This graph illustrates the proposed algorithm for selecting state space points for Monte Carlo integration using a minimal example. An explanation is given in the text.

To illustrate this algorithm, consider how it might be applied to a simple dynamic labour force participation model in the spirit of Eckstein and Wolpin (1989) with $T = 5$ periods and a binary choice (so $K = 2$) between market work ('W') and home production ('H'). Suppose

[10] Like the systematic sampling algorithm, this algorithm was primarily chosen for its simplicity; it is unlikely to be the optimal choice.

there is only one state variable that persists across periods: work experience. The stock of work experience is equal to the number of periods spent in market work. The econometrician observes a balanced panel of choices $d_{it} \in \{W, H\}$ for each individual i and time period $t \leq T$.

Figure 1 maps out the state space for this model. There are a total of 15 state space points: in each period t, the stock of work experience up to t can take t different values between $x_t = 0$ (market work in no previous period) and $x_t = t - 1$ (market work in every previous period). Agents can move between state space points as indicated by the arrows. The expected value of being at each state space point can be obtained by backward recursion as in the canonical KW94 model.

Now suppose an approximate solution of this model were to be obtained using the approximation method of KW94, with a maximum of $M = 3$ state space points per period evaluated using Monte Carlo integration, and the rest using an interpolation procedure as outlined. Assume further that in fact most agents work in most periods, so state space points towards the bottom of Figure 1 are reached much more frequently than those near the top, as indicated by the outline thickness. In particular, suppose that all agents in the data have gained at least one year of work experience when they reach period $t = 4$, so the state space points where $x_3 = 0$ and $x_4 = 0$ are never reached (indicated by dashed lines).

Our data-driven selection scheme ensures that nodes that are commonly reached (such as those along the bottom edge of Figure 1) are evaluated with high probability. The intuition behind this procedure is that higher precision is crucial for nodes that are on or near the optimal path for most agents, but less important for those that are not. By evaluating state space points along 'popular' paths, our selection scheme thus achieves a more efficient allocation of computational resources.

The shading in Figure 1 marks the state space points that might be selected for Monte Carlo integration by our procedure in this example. In periods $t = 2$ and $t = 3$, all state space points are evaluated, as $M = 3$ is larger than or equal to the total number of state space points that can be reached.[11] In period $t = 4$, all nodes that are reached in the data will be evaluated. In period $t = 5$, the number of state space points reached in the data is greater than M, so the outcome of the selection algorithm depends on the random ordering of individuals in step 1; one possible outcome is shown.

[11] The expected value of $x_0 = 0$ need not be evaluated, as it is irrelevant for agents' choices.

3.3 Improvement 3: no interpolation when multicollinearity is detected

As the regressors in the interpolation regression (11) are (square roots of) the differences between the option with the highest expected value and the expected value of each option, the value of the regressors relating to the option with the highest expected value is in each case zero. If in a given period one option nearly always has the highest expected value, regardless of an agent's previous path through the model, the interpolation regression will therefore suffer from a multicollinearity problem. This problem is particularly acute in models with substantial *ex ante* heterogeneity, as an agent's type in these models may to a large extent determine their path through the model.

As a result, the moment matrix $\boldsymbol{X}'\boldsymbol{X}$ can be singular or nearly singular. In the former case, the matrix will not be invertible, so it will not be possible to perform the interpolation regression at all. In the latter case, interpolation will be possible, but the parameters will be very sensitive to the values for the small number of nodes where the dominant option does not have the highest expected value.

For the approximation results below, we always set $g(.) = 0$ when the moment matrix $\boldsymbol{X}'\boldsymbol{X}$ is singular. However, this leaves two problems. First, in cases where $\boldsymbol{X}'\boldsymbol{X}$ is nearly singular, the approximation can still go badly awry, resulting in a low proportion of correct choices and very poor fit, as approximation errors in one period affect choices in all earlier periods. Second, dispensing with the interpolation regression when $\boldsymbol{X}'\boldsymbol{X}$ is singular creates discontinuous jumps in the objective surface at the boundary between singularity and near-singularity, potentially impairing estimation. Our third suggestion for applied researchers, therefore, is to set $g(.) = 0$ in cases where $\boldsymbol{X}'\boldsymbol{X}$ is nearly singular as well, and smoothly interpolate between that case and the regular interpolation regression.[12]

One way to accomplish this is to use the (1-norm) *condition number* for inversion of $\boldsymbol{X}'\boldsymbol{X}$, $\kappa(\boldsymbol{X}'\boldsymbol{X})$. We call $\boldsymbol{X}'\boldsymbol{X}$ "nearly singular" if and only if $\kappa(\boldsymbol{X}'\boldsymbol{X}) > 1/\epsilon$, where $\epsilon = 2^{-52} \approx 2.2 \times 10^{-16}$ is the machine epsilon for floating point arithmetic on 64-bit computer systems. For smooth interpolation, we define the polynomial smoothing kernel:

[12] An alternative approach would be to regularize the least squares problem using ridge regression (Hoerl and Kennard, 1970).

$$K(v) = \begin{cases} 0 & \text{for } v < 0 \\ \frac{v^2}{2\phi^2} & \text{for } 0 \leq v < \phi \\ -\frac{v^2}{2\phi^2} + \frac{2v}{\phi} - 1 & \text{for } \phi \leq v < 2\phi \\ 1 & \text{for } v \geq 2\phi \end{cases} \tag{12}$$

where ϕ is a smoothing parameter that we set to $\phi = 50$.[13] We then set $v = \frac{1}{\epsilon\kappa(\boldsymbol{X}'\boldsymbol{X})} - 1$ and replace (9) by

$$\underbrace{\tilde{\mathrm{E}}[\max_{k \leq K} V_{kt}(\boldsymbol{S}_t)]}_{\widetilde{\mathrm{EMAX}}} = \underbrace{\max_{k \leq K} \mathrm{E}\left[V_{kt}(\boldsymbol{S}_t)\right]}_{\mathrm{MAXE}} + K(v) g\left(\tilde{V}_{1t}(\boldsymbol{S}_t), \ldots, \tilde{V}_{Kt}(\boldsymbol{S}_t)\right). \tag{13}$$

3.4 Monte Carlo evidence: approximation performance

We evaluate the performance of different approximation methods using Monte Carlo simulations. All simulations are based on the same parameter values as the main results in KW94.[14] We first simulate 500 balanced panels of $N = 10,000$ individuals, recording their choices (and wages, in case a market work option is chosen) in each period. All of these simulations rely on the same 'true' model solution, obtained by fully solving the model by Monte Carlo integration using $D = 80,000$ simple random draws. We then compare these 500 simulated data sets to 500 data sets obtained using different approximation methods.

We use two different metrics to assess the performance of different methods. The first is the share of choices in the simulated data sets based on each approximation method that are the same as in the data sets simulated using the 'true' model solution, holding the sequence of realized shocks $(\boldsymbol{\epsilon}_{it})_{t=1}^{T}$ constant for each individual i. As realized shocks are held constant, any difference in agents' choices will be the result of approximation errors.

Our second metric is the weighted sum of squared deviations from key moments of the simulated data: the share of individuals choosing each option in each period, the mean log wage in each market occupation in each period and the variance of log wages in each occupation in each period.[15] The weight matrix is the inverse of the diagonal of the variance matrix

[13] A similar polynomial kernel is recommended in Hajivassiliou, McFadden and Ruud (1996).

[14] More specifically, these are the parameter values corresponding to their 'data set one'. For individual parameter values, see Appendix E of this paper or Table 1 of KW94.

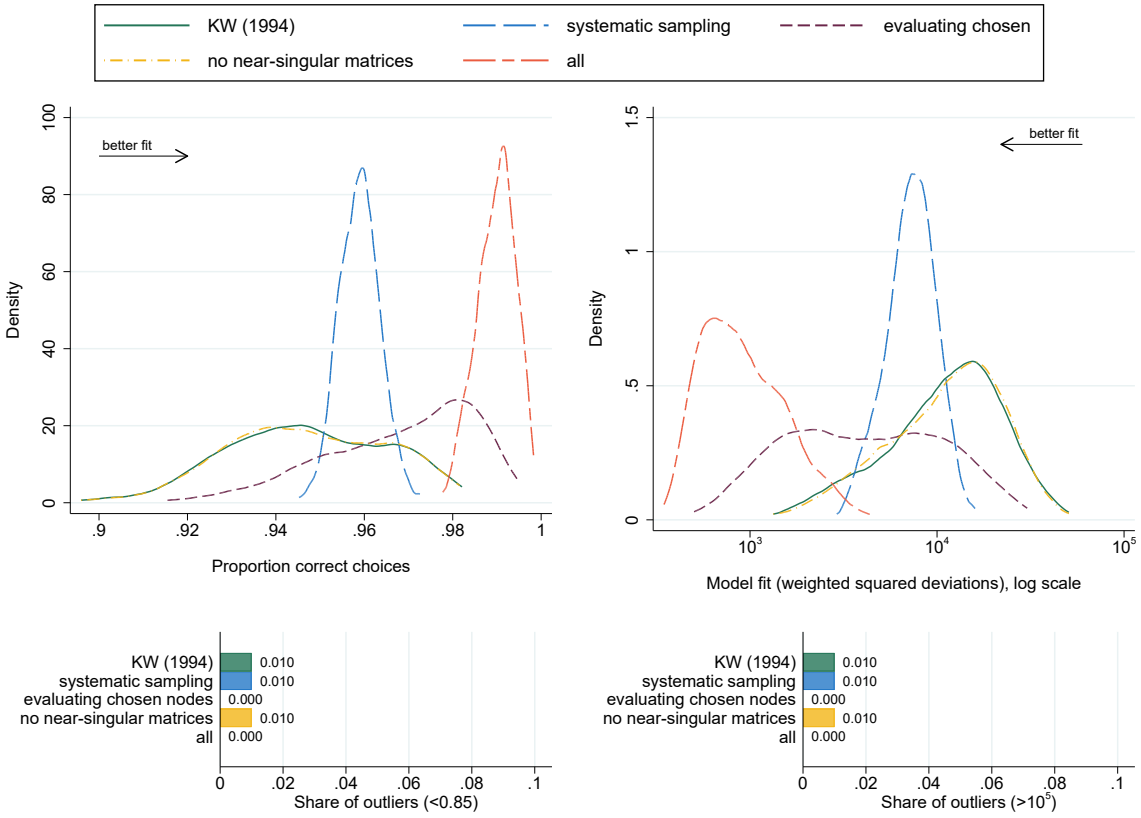
[15] Some choice shares are excluded to avoid linear dependence.

of the vector of data moments. Formally, our measure of model fit φ is given by

$$\varphi(\boldsymbol{\theta}) = (\boldsymbol{m} - \hat{\boldsymbol{m}})'(\text{diag}(\hat{\boldsymbol{V}}))^{-1}(\boldsymbol{m} - \hat{\boldsymbol{m}}) \tag{14}$$

where $\boldsymbol{m}$ is the vector of moments and $\hat{\boldsymbol{V}}$ is the sample variance matrix of $\boldsymbol{m}$.[16]

Figure 2: Approximation performance of different versions of the Keane and Wolpin method



Notes. The densities shown are kernel densities estimated using an Epanechnikov kernel with the bandwidth selected using Silverman's Rule. In each case, the kernel density is calculated over 500 approximation runs with different shocks and approximation draws. In all cases, a maximum of $M = 500$ nodes were evaluated each period, $D = 2000$ draws were used for Monte Carlo integration, and $N = 10,000$ individuals' choices were simulated. The two panels show our two different metrics of approximation performance. Realized shocks $(\boldsymbol{\epsilon}_{it})_{t=1}^{T}$ are held constant only for the evaluation of the proportion of correct choices (left panel).

For this second metric, not only the approximation draws but also the realized shocks $(\boldsymbol{\epsilon}_{it})_{t=1}^{T}$ are allowed to vary. While this measure is less straightforward to interpret than the share of correct choices, it is a more dependable metric of the distance in simulation outcomes between the true model and different approximations, as it does not rely on knowledge of

[16] This weight matrix is commonly used in empirical work. Examples include Blundell et al. (2016), Morten (2019) and Blundell et al. (2021).

unobservables (the realized shocks $(\epsilon_{it})_{t=1}^{T}$).[17]

Figure 2 shows how our amendments improve the KW94 method when a moderate number of state space points ($M = 500$) is evaluated each period. In all cases, $D = 2000$ draws were used for Monte Carlo integration, and $N = 10,000$ individuals' choices were simulated. The two panels show our two metrics of approximation performance.

The traditional KW94 method performs well on the whole, but the fit varies substantially across approximations run. The share of 'correct' choices varies roughly between 90 percent and 98 percent (disregarding outliers). The model fit measured by weighted squared deviations also varies substantially between approximation runs (more than an order of magnitude), even if outliers are disregarded.

Both the use of systematic sampling and the evaluation of chosen nodes substantially improve performance on their own. Our third improvement has little effect on approximation performance in the KW94 model at the true parameters, as (near-)singular moment matrices $\boldsymbol{X}'\boldsymbol{X}$ typically only occur in periods $t = 2$ and $t = 3$ when the number of possible state space points is low enough for all expectations to be evaluated by Monte Carlo integration. The combination of all amendments is much better still. For the variation combining all changes, the proportion of correct choices is always above 97 percent. The median weighted sum of squared deviations from true moment values is roughly an order of magnitude lower in the variation with all changes than using the traditional KW94 method.

Figure 3 shows how our changes to the KW94 method affect the trade-off between accuracy and computation time. For each series, the four data points shown represent, from left to right, the KW94 method with $M = 250$, $M = 500$ and $M = 2000$ evaluation points, and the full solution. Again, the two panels show our two metrics of approximation performance.[18]

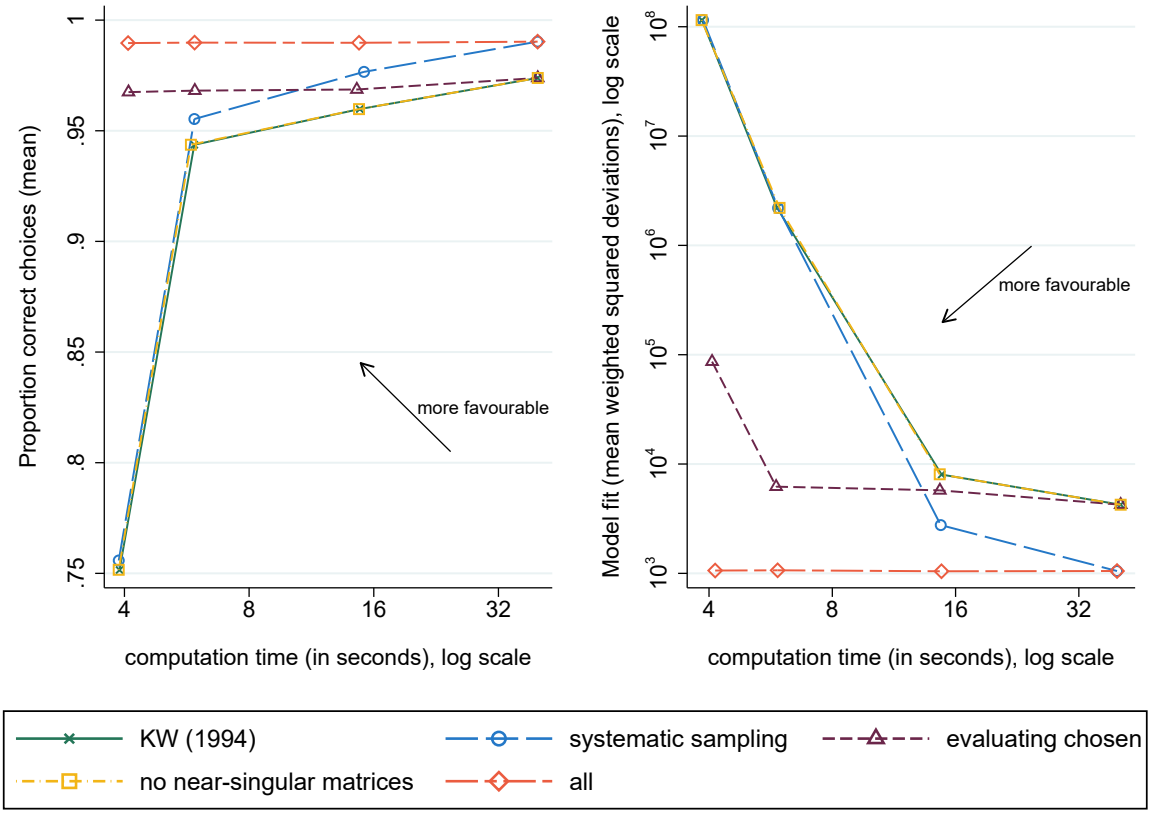
Our improvements fundamentally change the nature of the trade-off between accuracy and computation time for the KW94 method. In the traditional version, the trade-off is quite steep even with a large number of nodes evaluated using Monte Carlo simulation. In contrast, when evaluation nodes are selected systematically, the trade-off essentially disappears for moderate numbers of nodes evaluated by Monte Carlo integration. With all improvements, excellent approximation performance is achieved even when at most $M = 250$ nodes are evaluated each

[17] Importantly, using this second metric, we can assess whether the advantage of evaluating nodes chosen in the data depends on assuming the same realized shocks.

[18] The apparent difference in the shape of the trade-off across the two measures is mainly a result of the higher sensitivity to outliers of the average sum of weighted squared deviations compared to the average share of correct choices.

period.

Figure 3: Trade-off between accuracy and computation time



Notes. In each case, the mean over 500 approximation runs with different shocks and approximation draws is shown. In all cases, $D = 2000$ draws were used for Monte Carlo integration, and $N = 10,000$ individuals' choices were simulated. For each series, the four data points shown represent, from left to right, the KW94 method with $M = 250$, $M = 500$ and $M = 2000$ evaluation points, and the full solution. The two panels show our two different metrics of approximation performance. Realized shocks $(\boldsymbol{\epsilon}_{it})_{t=1}^{T}$ are held constant only for the evaluation of the proportion of correct choices (left panel).

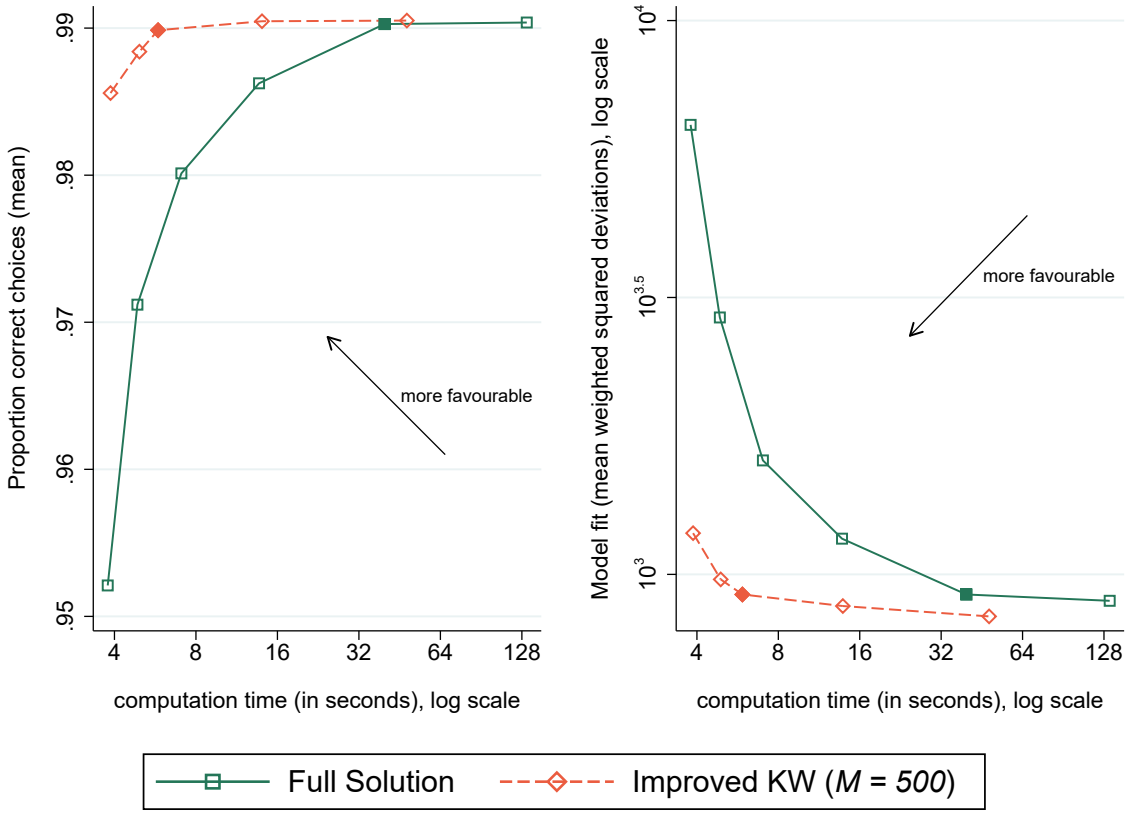
Figure 4 compares our improved KW94 method with $M = 500$ evaluation points to the full solution, with the number of draws adjusted to achieve comparable computation times. At nearly all levels of computational expense, the improved KW94 method delivers superior approximation accuracy, with a larger lead at shorter computation times. The full solution reaches similar accuracy only at computation times larger than around one minute.

4 Estimation of the canonical model

This section shows that these gains in approximation performance for a known set of parameters translate into similar gains in estimation performance when the parameters are unknown

and need to be estimated from observed data. Using the improved KW94 method, the parameters of the canonical model can be estimated with a high degree of accuracy even when only a small number of nodes are evaluated using Monte Carlo integration. This in turn translates into more accurate predictions in a simple policy experiment.

Figure 4: Trade-off between accuracy and computation time: comparison to full solution



Notes. In each case, the mean over 500 approximation runs with different shocks and approximation draws is shown. In all cases, $N = 10,000$ individuals' choices were simulated. For the full solution, the six data points shown represent, from left to right, the solution with $D = 50$, $D = 100$, $D = 200$, $D = 500$, $D = 2000$ and $D = 5000$ draws. For the improved KW94 method, the five data points shown represent, from left to right, the solution with $D = 500$, $D = 1000$, $D = 2000$, $D = 5000$, and $D = 20,000$ draws. Filled symbols indicate $D = 2000$, the number of Monte Carlo draws used in other graphs. The two panels show our two different metrics of approximation performance. Realized shocks $(\boldsymbol{\epsilon}_{it})_{t=1}^{T}$ are held constant only for the evaluation of the proportion of correct choices (left panel). For comparability, systematic sampling was used for both methods.

These results obtain regardless of whether the model is estimated using Simulated Maximum Likelihood as in the original KW94 contribution and much of the literature at the time, or using a simple Simulated Method of Moments or Indirect Inference method, as has become more popular recently. Simulated Maximum Likelihood can provide more accurate results, but this typically comes at the cost of higher computation time, as numerical integration is

required to evaluate the likelihood function.[19] We present Simulated Maximum Likelihood results below; analogous Simulated Method of Moments results are presented in Appendix A.

As in KW94, all of the results in this section are presented for a fixed (maximum) number of nodes evaluated by Monte Carlo integration, a fixed number of numerical integration draws for both solution and estimation, and with the smoothing parameter held constant. The motivation for this is that it makes our method more transparent and the results easier to interpret. In actual estimation, it would be advisable to increase the number of nodes evaluated by Monte Carlo integration, increase the number of integration draws, and reduce the amount of smoothing as the algorithm converges.[20]

In order to avoid spuriously positive results, we start the estimation procedures not at the true values but at a set of values that an empirical researcher without knowledge of the true parameter values might reasonably have picked as starting values. In particular, for the wage parameters, we use the parameters of two separate bivariate sample selection models estimated from the data.[21] We further set the value of home production γ_0 to one standard deviation below the median blue collar wage, and the standard deviation of the non-wage options to the average estimated standard deviation of the market-work options. All other parameters are set to zero.

Our Simulated Maximum Likelihood procedure closely follows KW94. The log-likelihood function takes the form

$$l(\theta) = \sum_{i=1}^{N} \left[\log f(\boldsymbol{d}_{i1}, w_{i1}) + \sum_{t=2}^{T} \log f(\boldsymbol{d}_{it}, w_{it} | \{\boldsymbol{d}_{i,s}\}_{s=1}^{s=t-1}) \right]. \tag{15}$$

where $f(\boldsymbol{d}_{it}, w_{it})$ is the joint density of agent i's choice in period t and the corresponding observed wage. If no wage is observed, because the agent chose schooling ($k = 3$) or home production ($k = 4$), $f(\boldsymbol{d}_{it}, w_{it})$ is simply the probability of the observed choice $P(\boldsymbol{d}_{it})$ conditional

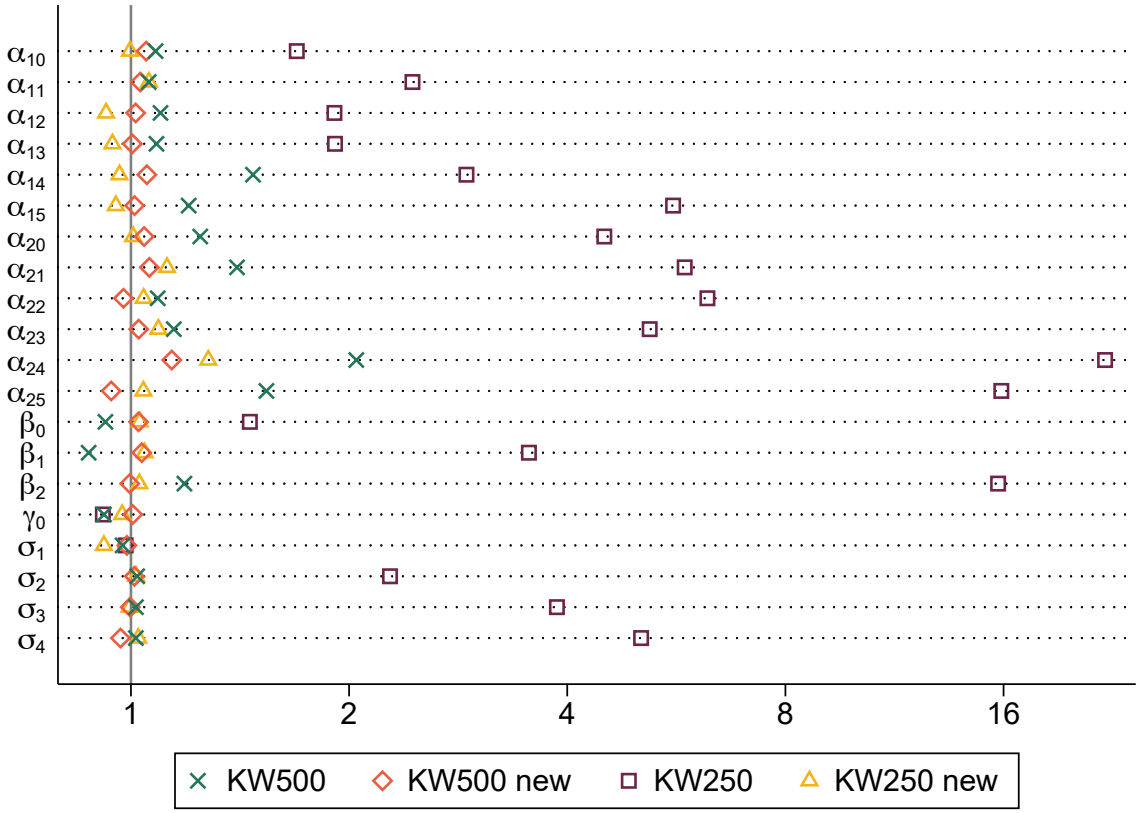
[19] See Eisenhauer, Heckman and Mosso (2015) for a comparison of the two approaches in the context of dynamic discrete choice models.

[20] One way of implementing this would be as a two-step procedure, where (at least) one final step of a (Quasi-)Newton algorithm would be performed using the full solution, a higher number of draws, and less smoothing. This would likely yield additional accuracy gains at little computational cost. For theoretical arguments in favour of this two-step approach, see section 3.1 of Hajivassiliou (2000) and Kristensen and Salanié (2017). Kristensen and Salanié (2017) and Bruins et al. (2018) present Monte Carlo evidence suggesting that this approach also works well in practice.

[21] Two natural exclusion restrictions arise from the full model: it is known that which period an individual is in and whether school was chosen in the previous period can both affect continuation values and thus choices but not wages. We estimated the selection models using Maximum Likelihood (assuming normally distributed errors). Using the Heckman (1979) two-step method would have resulted in starting values at a similar distance from the the true parameters.

on past choices.

Figure 5: Root mean square error of parameter estimates relative to full solution



Notes. The diagram shows the relative root mean squared error for all model parameters. Root mean squared errors obtained using the full solution are normalized to unity as indicated by the grey vertical line. For each approximation method, root mean squared errors were calculated over 24 sets of estimated parameters. These were obtained by estimating the model parameters 30 times from different simulated data sets; the 6 estimated parameter sets (20%) with the lowest simulated likelihood were dropped to guard against the effects of outliers and numerical problems. Each simulated data set used for estimation was a balanced panel of $N = 2,000$ individuals. $D = 2000$ solution draws were used in model solution. For each estimation run, 15 steps of the BHHH algorithm were performed. For each observation, 200 Halton draws were used to simulate the likelihood. Likelihood simulation draws were held constant across estimation runs in order to minimize statistical noise unrelated to the different approximation methods.

We simulate the density $f(\boldsymbol{d}_{it}, w_{it})$ using the simulation technique proposed in chapter 1 of this thesis, which in the absence of measurement error is numerically equivalent to the procedure likely used in KW94.[22] The probabilities that enter the likelihood function are smoothed using the smoothed-Logit simulator first proposed by McFadden (1989).[23] This is necessary to avoid numerical problems related to simulated probabilities of zero; it has the important added advantage that it leads to a smooth log-likelihood function and thus permits the use of

[22] See chapter 1 for details.
[23] Like KW94, we use a tuning parameter of 500.

fast and easily parallelized gradient-based optimization routines.[24]

Estimation performance. Figure 5 shows the root mean squared error of all estimated model parameters, relative to the root mean squared error when the full solution with the same number of draws is used. Thirty estimation runs were performed with different simulation draws for the model solution, but with the likelihood simulation draws held constant in order to minimize statistical noise unrelated to the different approximation methods. The six estimation runs (20 percent) with the lowest simulated likelihood at the estimated parameters were dropped to guard against outliers and numerical problems.[25] In each case, the parameters were estimated from a balanced panel of $N = 2000$ people that was simulated from the 'true' model. $D = 2000$ solution draws were used in model solution and each density $f(\boldsymbol{d}_{it}, w_{it})$ was simulated using 200 Halton draws.[26] For each estimation run, 15 steps of the BHHH algorithm were performed; further iterations were found not to change estimated parameter values substantially regardless of the solution method. The root mean squared error was chosen as a measure of accuracy as it captures both bias in the estimated parameter values and their variability.[27]

The root mean squared errors for the improved version of the KW94 method are very close to the root mean squared errors for the full solution, even when at most $M = 250$ state space points are evaluated per period. Root mean squared errors are mostly slightly higher for the traditional version of the KW94 method with $M = 500$ and much higher with $M = 250$.

Policy experiment. In order to gauge the economic significance of these differences in estimation performance, we follow KW94 and perform a simple policy experiment: we investigate the impact of a \$500 tuition subsidy on time spent in each occupation. We simulate two balanced panels from the model for each of the 24 sets of estimated parameters (with and without the subsidy). Each data set contained $N = 10,000$ individuals, and $D = 2,000$ draws were used in Monte Carlo Integration. We used new draws for both realized shocks and numerical

[24] We use the BHHH algorithm (Berndt et al., 1974), as extensive experimentation has shown this algorithm to be much more efficient than other commonly used algorithms such as Simplex, Simulated Annealing, or BFGS.

[25] Estimation failures, marked by a numerical derivative of zero or estimated parameter values implausibly far away from the starting values, were assigned a likelihood of negative infinity (the number of such failures was never higher than six, the number of dropped estimation runs).

[26] The same draws were used for all individuals, and only varied across period t and option k. Halton draws were used to maximize accuracy at a given computational cost.

[27] The *relative* rather than *absolute* root mean squared error is shown, as the parameters have very different scales. For separate results for bias and standard deviation (comparable with Keane and Wolpin, 1994), see Appendix E.

integration. In all cases, simulations were performed using the full solution method.

Table 1: $500 tuition fee subsidy: Simulated Maximum Likelihood

	Truth	Starting	Full	KW250	KW250 New	KW500	KW500 New
Blue Collar	-3.34	-.126	-3.621	-2.31	-3.596	-3.575	-3.595
	(.119)	(.010)	(.193)	(1.652)	(.213)	(.359)	(.180)
White Collar	2.079	-.235	2.282	1.387	2.257	2.376	2.256
	(.109)	(.013)	(.180)	(1.291)	(.194)	(.402)	(.170)
School	1.461	.408	1.535	1.061	1.532	1.399	1.533
	(.026)	(.005)	(.042)	(.456)	(.045)	(.043)	(.037)
Home	-.199	-.048	-.195	-.138	-.192	-.199	-.194
	(.011)	(.003)	(.015)	(.082)	(.017)	(.032)	(.014)

Notes. Estimated impact of a $500 tuition subsidy on average years spent in each occupation. Two balanced panels were simulated from the model for each of 24 sets of estimated parameters (with and without the subsidy). Sample standard deviations are given in parenthesis. Each data set contains $N = 10,000$ individuals, and $D = 2,000$ draws are used in Monte Carlo Integration. Both realized shocks and draws for numerical integration were different from those used in estimation. In all cases, simulations were performed using the full solution method.

Table 1 shows the estimated effect of the subsidy on the average time spent in each occupation. At the true parameters, the subsidy leads to a rise in the average number of years individuals spend in education and white-collar work, and a fall in the number of years spent in blue-collar work and home production. The column labelled 'Starting' shows the simulated effect when the starting parameters are used instead; even though these parameters are in the vicinity of the true parameters, the simulated effect is very different compared to the simulated effect when the true parameters are used, underlining the importance of accurately estimated parameters.

Simulation based on parameters estimated using the full solution comes close to simulation results obtained using the true parameters, but the magnitude of the effect is slightly overestimated on average. Using the traditional KW94 approximation method with $M = 500$ evaluated nodes and the improved method with $M = 250$ and $M = 500$ nodes, similar simulation values are obtained on average. However, crucially, the improved method leads to much lower standard deviations of the simulated effect. The difference is especially dramatic when at most $M = 250$ nodes are evaluated by Monte Carlo integration each period: in that case, the improved method still comes close to the results using the full solution, whereas the traditional method yields very inaccurate results, echoing the findings for the root mean squared error of the parameter estimates.

5 A model with unobserved heterogeneity

This section introduces unobserved *ex ante* heterogeneity across individuals into the KW94 model and evaluates the performance of both the traditional KW94 approximation method and our improved version in that setting. The motivation is that such heterogeneity is a standard assumption of nearly all applied work in labour economics that uses dynamic discrete choice models.[28] As shown below, the extent of *ex ante* heterogeneity can have a large impact on the performance of different approximation methods. In particular, the traditional method of KW94 can be unreliable when *ex ante* heterogeneity is large. This has likely been a key reason for the recent decline in the use of this method among applied researchers.

We modify the canonical model of KW94 to allow for two different types of agent, where the type is known to the agent but not to the econometrician. An agent's type influences their payoffs from market work as follows:

$$u_{1t} = w_{1t} = \exp(\alpha_{10} + \zeta_1 1(\tau_i = 1) + \alpha_{11} s_t + \alpha_{12} x_{1t} - \alpha_{13} x_{1t}^2 + \alpha_{14} x_{2t} - \alpha_{15} x_{2t}^2 + \epsilon_{1t}) \tag{16}$$

$$u_{2t} = w_{2t} = \exp(\alpha_{20} + \zeta_2 1(\tau_i = 2) + \alpha_{21} s_t + \alpha_{22} x_{2t} - \alpha_{23} x_{2t}^2 + \alpha_{24} x_{1t} - \alpha_{25} x_{1t}^2 + \epsilon_{2t}) \tag{17}$$

where τ_i is agent i's type and $1(.)$ is the indicator function. We set $\zeta_1 = 0.05$ and $\zeta_2 = 0.1$; all other parameters are unchanged. This implies that type 1 agents enjoy a wage premium of roughly 5% in blue-collar work, and type 2 agents enjoy a wage premium of around 10% in white-collar work. The share of each type is held fixed at one half.[29]

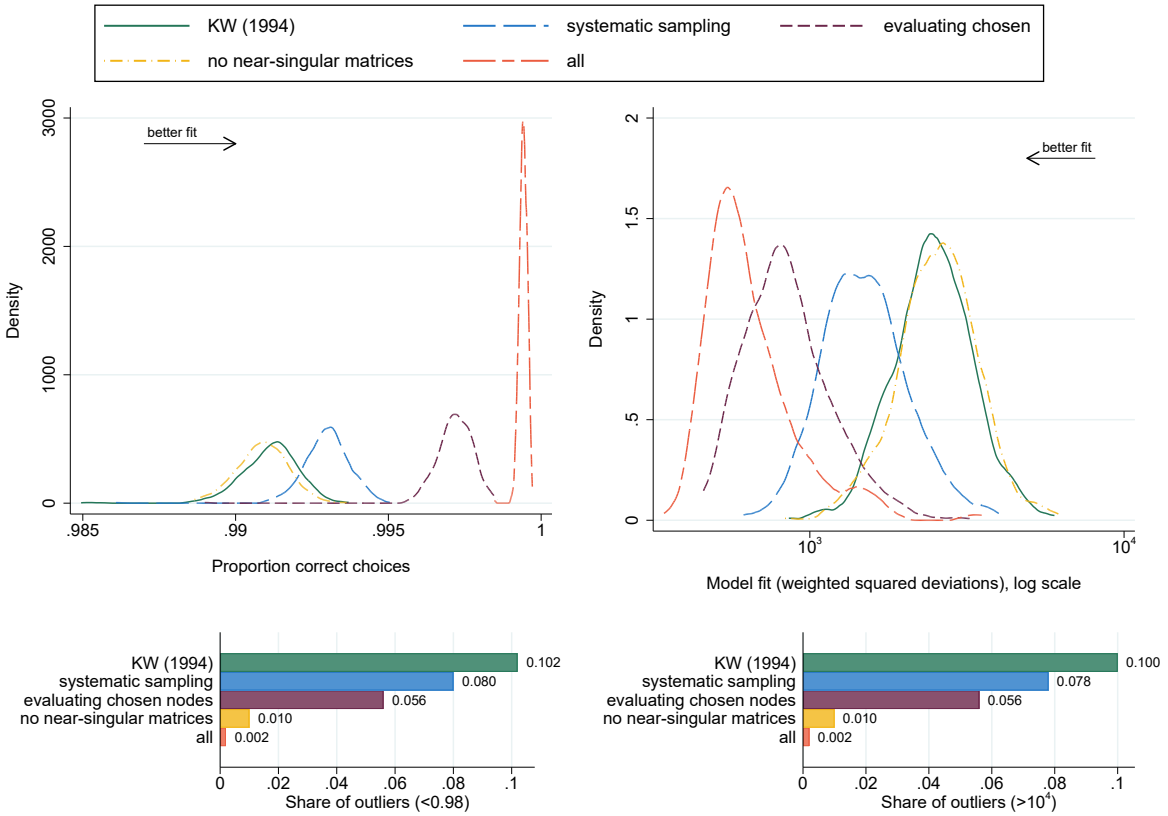
5.1 Approximate solution

One difference arising from the introduction of unobserved heterogeneity into the model is that the state space is now twice as large, as the model now needs to be solved separately for each type. This roughly doubles the computation time required for solving the model regardless of which method is used. As discussed in section 3.3, a particular challenge for the KW94 method in models with *ex ante* heterogeneity is that multicollinearity problems in the interpolation regression are especially likely to occur.

[28] Notably, this includes Keane and Wolpin (1997), which first applied the KW94 method in empirical work.

[29] We fix the type share here to avoid identification challenges that are unrelated to the argument in this paper. When simulating data from the model, we assigned types to each individual according to $\tau_i = 1 + 1(u_i > 1/2)$, where u_i is drawn from the standard uniform distribution.

Figure 6: Approximation performance: model with heterogeneity



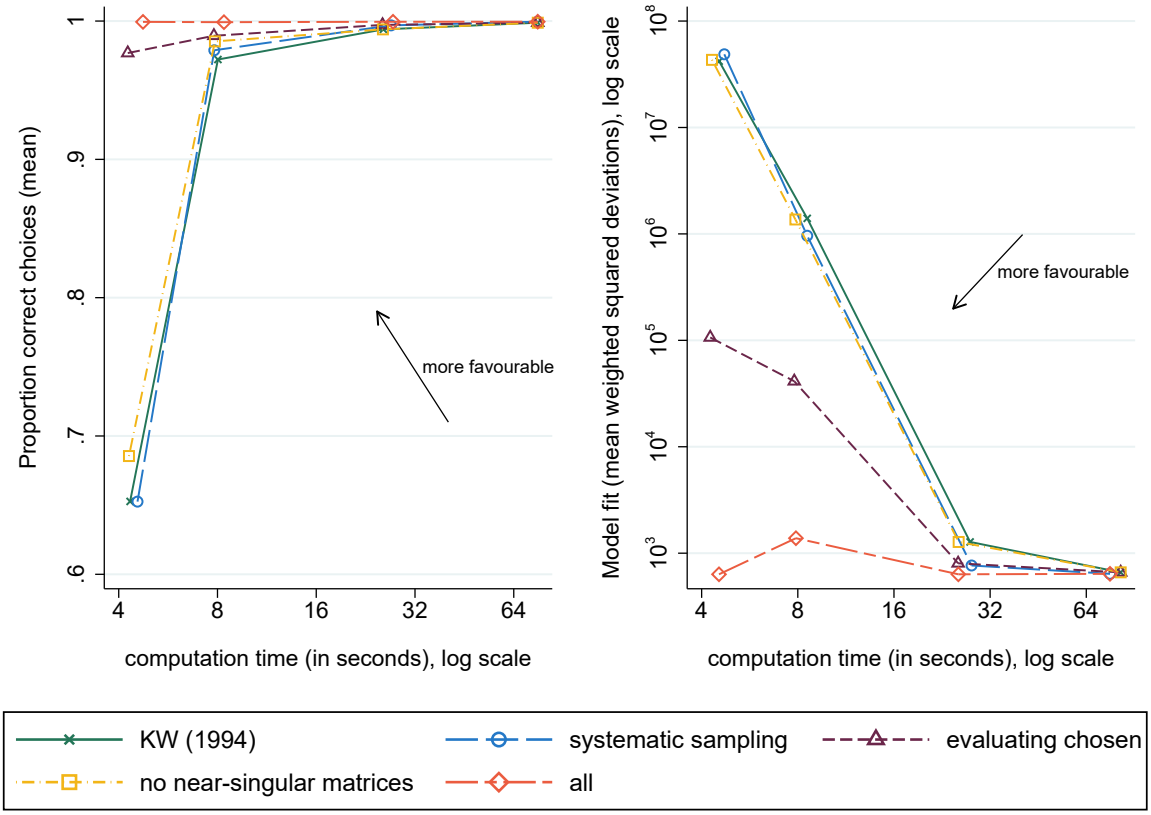
Notes. The densities shown are kernel densities estimated using an Epanechnikov kernel with the bandwidth selected using Silverman's Rule. In each case, the kernel density was calculated over 500 approximation runs with different shocks and approximation draws. In all cases, a maximum of $M = 500$ nodes were evaluated each period, $D = 2000$ draws were used for Monte Carlo integration, and $N = 10,000$ individuals' choices were simulated. The two panels show our two different metrics of approximation performance. Realized shocks $(\boldsymbol{\epsilon}_{it})_{t=1}^{T}$ and types are held constant only for the evaluation of the proportion of correct choices (left panel).

We evaluate the performance of the different varieties of the KW94 method using the same metrics as above. Figure 6 shows the results when at most $M = 500$ nodes are evaluated by Monte Carlo integration. On the whole, the KW94 method works exceptionally well in the model with unobserved heterogeneity. Even using the basic method, around 99 percent of choices are the same in the approximately solved model in most cases. However, in a minority of cases, the basic KW94 approximation — as well as the variations with systematic sampling and data-guided node selection — can go badly awry, resulting in a low proportion of correct choices and very poor fit.

Virtually all of these 'approximation failures' appear to be due to multicollinearity in the interpolation regression. Dispensing with interpolation when the moment matrix $\boldsymbol{X}'\boldsymbol{X}$ is nearly singular removes almost all of them. The version of the KW94 method incorporating all

three improvements achieves excellent fit and a near perfect proportion of correctly predicted choices. The explanation for this extremely good performance — even compared to the same method in the model without heterogeneity — is that in the model with heterogeneity, fewer individuals are on the margin between different options, so the heterogeneity to a large extent 'pre-determines' agents' paths through the model.

Figure 7: Trade-off between accuracy and computation time: model with heterogeneity

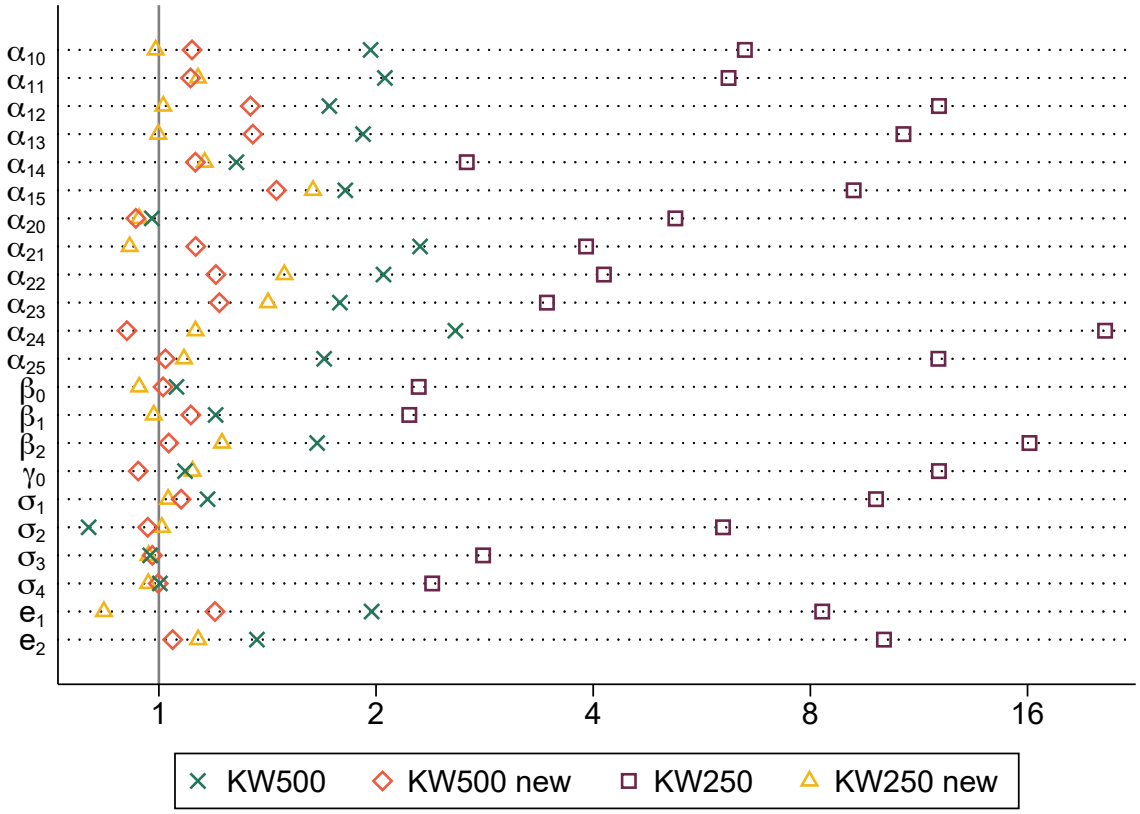


Notes. In each case, the mean over 500 approximation runs with different shocks and approximation draws is shown. In all cases, $D = 2000$ draws were used for Monte Carlo integration, and $N = 10,000$ individuals' choices were simulated. For each series, the four data points shown represent, from left to right, the KW94 method with $M = 250$, $M = 500$ and $M = 2000$ evaluation points, and the full solution. The two panels show our two different metrics of approximation performance. Realized shocks $(\boldsymbol{\epsilon}_{it})_{t=1}^{T}$ are held constant only for the evaluation of the proportion of correct choices (left panel).

Figure 7 shows how the mean fit in the model with heterogeneity varies as the number of points evaluated by Monte Carlo integration is increased. Notably, the full solution provides essentially perfect fit, whether or not systematic sampling is used for numerical integration. Evaluating the chosen nodes still substantially flattens the trade-off between accuracy and computation time. Again the trade-off virtually disappears (for the moderate numbers of

evaluation points shown) when all three improvements are combined.[30]

Figure 8: Root mean squared error of parameter estimates relative to full solution: model with heterogeneity



Notes. The diagram shows the relative root mean squared error for all model parameters. Root mean squared errors obtained using the full solution are normalized to unity as indicated by the grey vertical line. For each approximation method, root mean squared errors were calculated over 15 sets of estimated parameters. These were obtained by estimating the model parameters 30 times from different simulated data sets; the 15 estimated parameter sets (50%) with the lowest simulated likelihood were dropped to guard against the effects of outliers and numerical problems. Each simulated data set used for estimation was a balanced panel of $N = 2,000$ individuals. $D = 2000$ solution draws were used in model solution. For each estimation run, 30 steps of the BHHH algorithm were performed. For each observation, 200 Halton draws were used to simulate the likelihood. Likelihood simulation draws were held constant across estimation runs in order to minimize statistical noise unrelated to the different approximation methods.

5.2 Estimation

One preliminary challenge in estimating this model is that it is only set-identified (in a set of two elements). The reason is that for any given parameterization, a parameterization in which type 2 had the same payoffs as type 1 in the given parameterization and *vice versa* is empirically equivalent. In particular, it is easy to verify that any two parameterizations a and

[30]The slightly worse fit in terms of weighted squared deviations when at most $M = 500$ state space points are evaluated is due to the influence of a single outlier.

b are empirically equivalent if $\zeta_1^b = -\zeta_1^a$, $\zeta_2^b = -\zeta_2^a$, $\alpha_{10}^b = \alpha_{10}^a + \zeta_1^a$, $\alpha_{20}^b = \alpha_{20}^a + \zeta_2^a$, and all other parameters are identical. We address this challenge by reparameterizing the solution to the equivalent parameterization if and only if $\hat{\zeta}_1 < 0$ and $\hat{\zeta}_2 < 0$. The log likelihood function now takes the form

$$l(\theta) = \sum_{i=1}^{N} \log \sum_{j=1}^{J} P(\tau_i = j) f(\boldsymbol{d}_{i1}, w_{i1} | \tau_i = j) \prod_{t=2}^{T} f(\boldsymbol{d}_{it}, w_{it} | (\boldsymbol{d}_{i,s})_{s=1}^{s=t-1}, \tau_i = j). \tag{18}$$

Table 2: $500 tuition fee subsidy: model with heterogeneity

	Truth	Starting	Full	KW250	KW250 New	KW500	KW500 New
Blue Collar	-.141	-.009	-.138	-.374	-.141	-.147	-.137
	(.005)	(.001)	(.007)	(.517)	(.008)	(.013)	(.007)
White Collar	-.261	-.221	-.268	-.049	-.265	-.279	-.268
	(.006)	(.005)	(.007)	(.473)	(.006)	(.010)	(.010)
School	.403	.248	.406	.450	.406	.427	.405
	(.008)	(.005)	(.010)	(.107)	(.010)	(.013)	(.011)
Home	0	-.018	0	-.027	0	0	0
	(0)	(.001)	(0)	(.045)	(0)	(0)	(0)

Notes. Estimated impact of a $500 tuition subsidy on average years spent in each occupation. Two balanced panels are simulated from the model for each of 15 sets of estimated parameters (with and without the subsidy). Sample standard deviations are given in parenthesis. Each data set contains $N = 10,000$ individuals, and $D = 2,000$ draws are used in Monte Carlo Integration. Both realized shocks and draws for numerical integration are different from those used in estimation. Simulations are performed using the full solution method in all cases.

Figure 8 provides a measure of the performance of the different approximation methods when the model with heterogeneity is estimated using Simulated Maximum Likelihood. Estimation was performed exactly as for the model without heterogeneity, except that 30 (instead of 15) steps of the BHHH algorithm were performed and half (instead of 20%) of estimates were dropped to guard against outliers and convergence problems, reflecting the more complex numerical optimization problem.[31] As in the model without heterogeneity, our three improvements lead to more accurate parameter estimates that are roughly comparable to those obtained using the full solution. However, the results are noisier, which is partly explained by the lower number of estimation runs used in calculating the root mean squared error.

Table 2 again shows the impact of a $500 tuition subsidy. The relative performance of the traditional and the new versions of the KW94 method are similar to the canonical model. Notably, the true effect of the tuition subsidy is much smaller in the model with *ex ante* hetero-

[31] The starting vales for the additional model parameters ζ_1 and ζ_2 reflecting the heterogeneous payoffs for different types were set to zero.

geneity, even though the choice shares are similar. Again, the reason is that fewer individuals are on the margin between different options.

6 Conclusion

In this paper, we have suggested three improvements to the KW94 method, which until recently was commonly used by applied researchers to estimate finite-horizon discrete choice dynamic programming problems. First, systematic sampling substantially improves approximation performance. Second, drawing state space points for evaluation from the subset that agents reach in the data is much better than choosing them at random. Third, especially in models with *ex ante* heterogeneity, it is advisable to check for multicollinearity in the interpolation regression and dispense with polynomial interpolation when multicollinearity is detected.

With these improvements, the KW94 method achieves excellent approximation performance even in a model with a large state space and substantial *ex ante* heterogeneity. This suggests that the method has been abandoned prematurely. Its flaws are straightforward to address, and the potential savings in computation time make the effort worthwhile. Holding computation time constant, these improvements translate into more accurate parameter estimates and policy simulations; for a given error tolerance, a larger class of models can be estimated.

Future work is likely to improve on all three of our suggestions. More sophisticated systematic sampling techniques are likely to offer even larger variance reduction in Monte Carlo integration. Other systematic ways of choosing state space points to evaluate by numerical integration may improve on our approach. Other procedures for dealing with multicollinearity may also be superior to our technique; a particularly promising approach may be ridge regression (Hoerl and Kennard, 1970). Most importantly, it will be crucial to test our methods in other contexts besides the canonical KW94 model and its close cousins.

References

Abaluck, Jason, and Abi Adams-Prassl. 2021. "What Do Consumers Consider Before They Choose? Identification from Asymmetric Demand Responses." *The Quarterly Journal of Economics*. Forthcoming.

Altonji, Joseph G, Anthony A Smith Jr, and Ivan Vidangos. 2013. "Modeling earnings dynamics." *Econometrica*, 81(4): 1395–1454.

Arcidiacono, Peter, and Robert A Miller. 2011. "Conditional choice probability estimation of dynamic discrete choice models with unobserved heterogeneity." *Econometrica*, 79(6): 1823–1867.

Belzil, Christian, Jorgen Hansen, and Xingfei Liu. 2017. "Dynamic skill accumulation, education policies, and the return to schooling." *Quantitative Economics*, 8(3): 895–927.

Berndt, Ernst R, Bronwyn H Hall, Robert E Hall, and Jerry A Hausman. 1974. "Estimation and inference in nonlinear structural models." In *Annals of Economic and Social Measurement, Volume 3, number 4*. 653–665. NBER.

Blau, David M, and Donna B Gilleskie. 2006. "Health insurance and retirement of married couples." *Journal of Applied Econometrics*, 21(7): 935–953.

Blundell, Richard, Monica Costa Dias, Costas Meghir, and Jonathan Shaw. 2016. "Female labor supply, human capital, and welfare reform." *Econometrica*, 84(5): 1705–1753.

Blundell, Richard, Monica Costa-Dias, David Goll, and Costas Meghir. 2021. "Wages, Experience, and Training of Women over the Life Cycle." *Journal of Labor Economics*, 39(S1): S275–S315.

Bruins, Marianne, James A Duffy, Michael P Keane, and Anthony A Smith Jr. 2018. "Generalized indirect inference for discrete choice models." *Journal of Econometrics*, 205(1): 177–203.

Crawford, Gregory S, and Matthew Shum. 2005. "Uncertainty and learning in pharmaceutical demand." *Econometrica*, 73(4): 1137–1173.

Eckstein, Zvi, and Kenneth I Wolpin. 1989. "Dynamic labour force participation of married women and endogenous work experience." *The Review of Economic Studies*, 56(3): 375–390.

Eisenhauer, Philipp. 2019. "The approximate solution of finite-horizon discrete-choice dynamic programming models." *Journal of Applied Econometrics*, 34(1): 149–154.

Eisenhauer, Philipp, James J Heckman, and Stefano Mosso. 2015. "Estimation of dynamic discrete choice models by maximum likelihood and the simulated method of moments." *International Economic Review*, 56(2): 331–357.

Erdem, Tülin, and Michael P Keane. 1996. "Decision-making under uncertainty: Capturing dynamic brand choice processes in turbulent consumer goods markets." *Marketing Science*, 15(1): 1–20.

Geweke, John. 1996. "Monte Carlo simulation and numerical integration." *Handbook of Computational Economics*, 1: 731–800.

Geweke, John, and Michael Keane. 2000. "Bayesian inference for dynamic discrete choice models without the need for dynamic programming." *Simulation-based inference in econometrics: methods and applications*, 100–131.

Hajivassiliou, Vassilis. 2000. "Some practical issues in maximum simulated likelihood." *Simulation-Based inference in econometrics: Methods and Applications*, 71–99.

Hajivassiliou, Vassilis, Daniel McFadden, and Paul Ruud. 1996. "Simulation of multivariate normal rectangle probabilities and their derivatives theoretical and computational results." *Journal of Econometrics*, 72(1-2): 85–134.

Halton, John H. 1964. "Algorithm 247: Radical-inverse quasi-random point sequence." *Communications of the ACM*, 7(12): 701–702.

Heckman, James J. 1979. "Sample selection bias as a specification error." *Econometrica*, 153–161.

Hoerl, Arthur E, and Robert W Kennard. 1970. "Ridge regression: Biased estimation for nonorthogonal problems." *Technometrics*, 12(1): 55–67.

Hotz, V Joseph, and Robert A Miller. 1993. "Conditional choice probabilities and the estimation of dynamic models." *The Review of Economic Studies*, 60(3): 497–529.

Imai, Susumu, and Michael P Keane. 2004. "Intertemporal labor supply and human capital accumulation." *International Economic Review*, 45(2): 601–641.

Keane, Michael P, and Kenneth I Wolpin. 1994. "The solution and estimation of discrete choice dynamic programming models by simulation and interpolation: Monte Carlo evidence." *Review of Economics and Statistics*, 648–672.

Keane, Michael P, and Kenneth I Wolpin. 1997. "The career decisions of young men." *Journal of Political Economy*, 105(3): 473–522.

Keane, Michael P, and Kenneth I Wolpin. 2001. "The effect of parental transfers and borrowing constraints on educational attainment." *International Economic Review*, 42(4): 1051–1103.

Keane, Michael P, and Kenneth I Wolpin. 2010. "The role of labor and marriage markets, preference heterogeneity, and the welfare system in the life cycle decisions of black, hispanic, and white women." *International Economic Review*, 51(3): 851–892.

Kristensen, Dennis, and Bernard Salanié. 2017. "Higher-order properties of approximate estimators." *Journal of Econometrics*, 198(2): 189–208.

Kristensen, Dennis, Lars Nesheim, and Á de Paula. 2015. "CCP and the estimation of nonseparable dynamic models." Mimeo, University College London.

Lee, Donghoon, and Kenneth I Wolpin. 2006. "Intersectoral labor mobility and the growth of the service sector." *Econometrica*, 74(1): 1–46.

Lee, Donghoon, and Kenneth I Wolpin. 2010. "Accounting for wage and employment changes in the US from 1968–2000: A dynamic model of labor market equilibrium." *Journal of Econometrics*, 156(1): 68–85.

Llull, Joan. 2018. "Immigration, wages, and education: A labour market equilibrium structural model." *The Review of Economic Studies*, 85(3): 1852–1896.

McFadden, Daniel. 1989. "A Method of Simulated Moments for Estimation of Discrete Response Models Without Numerical Integration." *Econometrica*, 57(5): 995.

Morten, Melanie. 2019. "Temporary migration and endogenous risk sharing in village india." *Journal of Political Economy*, 127(1): 1–46.

Nadarajah, Saralees, and Samuel Kotz. 2008. "Exact distribution of the max/min of two Gaussian random variables." *IEEE Transactions on very large scale integration (VLSI) systems*, 16(2): 210–212.

Sauer, Robert M. 2015. "Does it pay for women to volunteer?" *International Economic Review*, 56(2): 537–564.

Skira, Meghan M. 2015. "Dynamic wage and employment effects of elder parent care." *International Economic Review*, 56(1): 63–93.

Stock, James H, and David A Wise. 1990. "Pensions, the Option Value of Work, and Retirement." *Econometrica*, 1151–1180.

Todd, Petra E, and Kenneth I Wolpin. 2006. "Assessing the impact of a school subsidy program in Mexico: Using a social experiment to validate a dynamic behavioral model of child schooling and fertility." *American Economic Review*, 96(5): 1384–1417.

Train, Kenneth E. 2009. *Discrete choice methods with simulation*. Cambridge University Press.

Ulyssea, Gabriel. 2018. "Firms, informality, and development: Theory and evidence from Brazil." *American Economic Review*, 108(8): 2015–47.

Van der Klaauw, Wilbert, and Kenneth I Wolpin. 2008. "Social security and the retirement and savings behavior of low-income households." *Journal of Econometrics*, 145(1-2): 21–42.

Appendix

A Canonical model: Simulated Method of Moments estimation

It is increasingly popular for applied researchers to estimate dynamic discrete choice models using the Simulated Method of Moments.[32] This can be advantageous, for example, when very large data sets are used, multiple data sets are combined, or evaluation of the likelihood is prohibitively complex. As demonstrated in this section, our results apply regardless of whether Simulated Maximum Likelihood or the Simulated Method of Moments is used.

We implement the Simulated Method of Moments approach using our second measure of model fit given in (14) as the criterion function.[33] Unlike in section 3.4, we do not simulate whole paths through the model, but in each period update the parameter vector with observed choices in the data. As a result, individuals in the simulated data set will in each period have the same vector of pre-determined state variables $\overline{s}_t$ as individuals in the 'true' data set, as the vector of pre-determined state variables in the canonical KW94 model only depends on observed choices. This approach has the computational advantage that small changes in model parameters typically only affect an individual's decision in a single period rather than in all subsequent periods, as future state variables are unaffected.[34] It is further useful to replace the simulated choices by smooth analogues in order to obtain an entirely smooth objective surface.[35] Then fast and easily parallelized gradient-based optimization routines can be used.[36]

Estimation performance. Figure 9 shows the estimation performance of different versions of the KW94 method relative to the full solution when the Simulated Method of Moments is used in estimation. Again estimation performance is measured by the root mean squared error relative to the full solution. Only the traditional KW94 method with $M = 250$ draws is clearly inferior to the full solution, with all other methods offering comparable performance.

[32] Important recent papers using this method include Altonji, Smith Jr and Vidangos (2013), Blundell et al. (2016), Llull (2018), Ulyssea (2018), Morten (2019) and Abaluck and Adams-Prassl (2021).

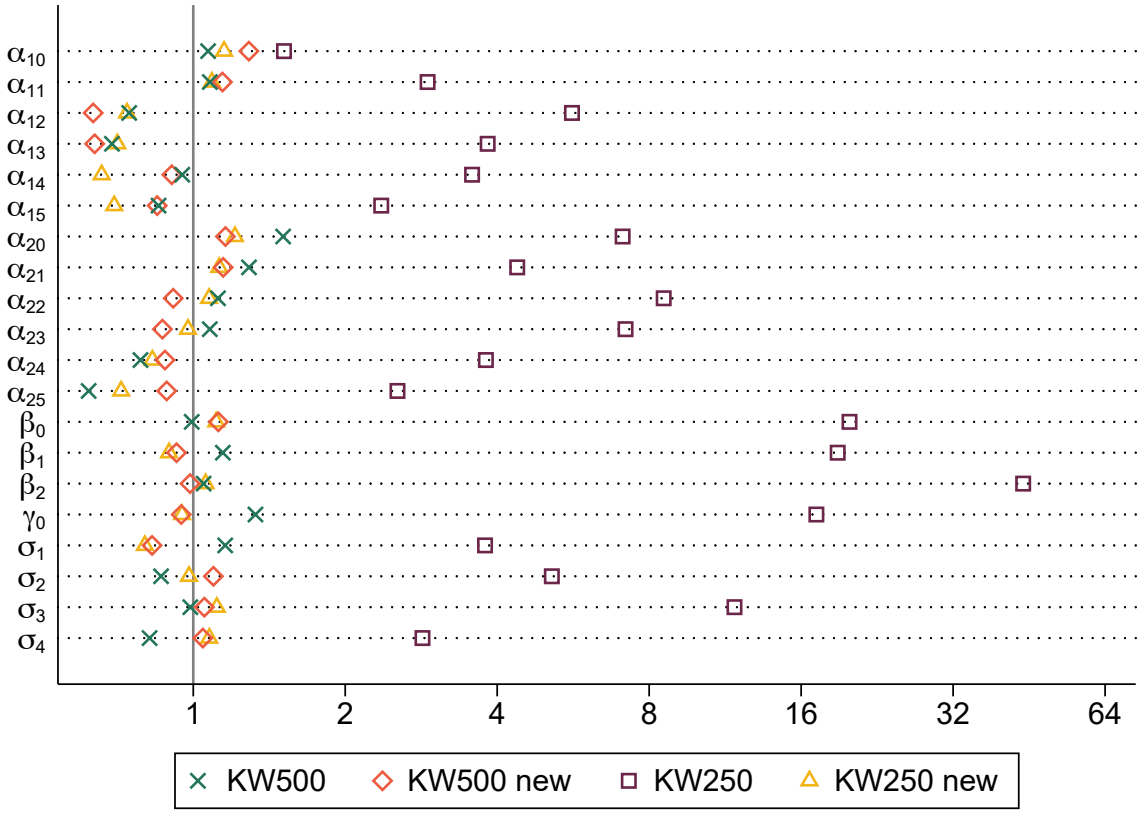
[33] In addition to the moments listed in section 3.4, we also include the entries of transition matrices between choices in each period as moments to aid identification.

[34] Skira (2015) takes a similar approach. An alternative would be to smooth over whole paths using the method from chapter 2 of this thesis.

[35] As shown by Bruins et al. (2018), the SMM estimator will still be consistent as the smoothing parameter goes to zero.

[36] We have found the Levenberg-Marquardt algorithm to be the most efficient.

Figure 9: SMM: root mean squared error relative to full solution



Notes. The diagram shows the relative root mean squared error for all model parameters. Root mean squared errors obtained using the full solution are normalized to unity as indicated by the grey vertical line. For each approximation method, root mean squared errors were calculated over 24 sets of estimated parameters. These were obtained by estimating the model parameters 30 times from different simulated data sets; the 6 estimated parameter sets (20%) with the lowest simulated likelihood were dropped to guard against the effects of outliers and numerical problems. Each simulated data set used for estimation was a balanced panel of $N = 10,000$ individuals. $D = 2000$ solution draws were used in model solution. For each estimation run, 15 steps of the Levenberg-Marquardt algorithm were performed.

Table 3: $500 Tuition Fee Subsidy: Simulated Method of Moments

	Truth	Starting	Full	KW250	KW250 New	KW500	KW500 New
Blue Collar	-3.34	-.126	-3.604	-2.624	-3.634	-3.58	-3.662
	(.119)	(.010)	(.178)	(1.315)	(.206)	(.314)	(.164)
White Collar	2.079	-.235	2.243	1.79	2.257	2.355	2.298
	(.109)	(.013)	(.173)	(.884)	(.181)	(.366)	(.166)
School	1.461	.408	1.552	.984	1.57	1.424	1.56
	(.026)	(.005)	(.054)	(.536)	(.061)	(.063)	(.034)
Home	-.199	-.048	-.191	-.150	-.193	-.199	-.195
	(.011)	(.003)	(.015)	(.087)	(.015)	(.027)	(.018)

Notes. Estimated impact of a $500 tuition subsidy on average years spent in each occupation. Two balanced panels are simulated from the model for each of 24 sets of estimated parameters (with and without the subsidy). Sample standard deviations are given in parenthesis. Each data set contains $N = 10,000$ individuals, and $D = 2,000$ draws are used in Monte Carlo Integration. Both realized shocks and draws for numerical integration are different from those used in estimation. In all cases, simulations were performed using the full solution method.

Policy experiment. Table 3 shows how these estimation results translate into performance in a simple policy experiment. As when parameters are estimated by Simulated Maximum Likelihood, simulation based on parameters estimated using the full solution comes close to simulation results obtained using the true parameters, although there is again a small bias that is very similar in magnitude and direction to the bias obtained when using Simulated Maximum Likelihood. The relative performance of the improved and traditional KW94 methods is similar to Simulated Maximum Likelihood.

B Further analytical bounds for the Keane and Wolpin method

While there is in general no analytical solution available for the continuation values in finite-horizon discrete choice dynamic programming models, it is generally possible to establish a lower bound. A basic lower bound that forms the basis of the KW94 approximation method is Jensen's Inequality. In the notation of section 3.4, imposing this bound is the same as requiring that $g(.) \geq 0$.

Further analytical bounds are often available. One avenue for deriving tighter bounds relies on the properties of the max function combined with Jensen's Inequality. For instance,

$$\begin{aligned} \mathrm{E}\left[\max(V_1, V_2, V_3, V_4)\right] &= \mathrm{E}\left[\max\left[\max(V_1, V_2), \max(V_3, V_4)\right]\right] \\ &\geq \max\left[\mathrm{E}\left[\max(V_1, V_2)\right], \mathrm{E}\left[\max(V_3, V_4)\right]\right]. \end{aligned} \tag{B.1}$$

The lower bound in the second line of (B.1) is weakly tighter than the bound provided by Jensen's inequality. This is helpful because it is often possible to derive an analytical form or a tighter bound for the expectation of the maximum of two options. In the canonical model of KW94, an analytical form is available for $\mathrm{E}\left[\max(V_3, V_4)\right]$ and a tighter bound for $\mathrm{E}\left[\max(V_1, V_2)\right]$ (taking future continuation values as given).

The analytical form for $\mathrm{E}\left[\max(V_3, V_4)\right]$ can be derived from the fact that V_3 and V_4 are jointly Normally distributed in the KW94 model. If $V_3 \sim N(\mu_3, \sigma_3^2)$, $V_4 \sim N(\mu_4, \sigma_4^2)$ and the correlation of V_3 and V_4 is ρ_{34}, then the expectation of the maximum of V_3 and V_4 takes the known form (see e.g. Nadarajah and Kotz, 2008):

$$\mathrm{E}\left[\max(V_3, V_4)\right] = \mu_3 \Phi\left(\frac{\mu_3 - \mu_4}{\theta}\right) + \mu_4 \Phi\left(\frac{\mu_3 - \mu_4}{\theta}\right) + \theta\phi\left(\frac{\mu_3 - \mu_4}{\theta}\right) \tag{B.2}$$

where $\theta = \sqrt{\sigma_3^2 + \sigma_4^2 - 2\rho_{34}\sigma_3\sigma_4}$ and Φ and ϕ are, respectively, the cdf and the pdf of the standard Normal distribution.

A tighter bound for $\mathrm{E}\left[\max(V_1, V_2)\right]$ can be derived from the fact that for $k = 1, 2$, V_{kt} can be decomposed into $V_{kt} = u_{kt} + \delta\,\mathrm{E}(V_{t+1}^k)$. Hence

$$\mathrm{E}\left[\max(V_{1t}, V_{2t})\right] \geq \mathrm{E}\left[\max(u_{1t}, u_{2t})\right] + \delta \min\left[\mathrm{E}(V_{t+1}^1), \mathrm{E}(V_{t+1}^2)\right]. \tag{B.3}$$

As there is no continuation value in the final period, the second term drops out and the expression holds with equality if $t = T$.

u_{1t} and u_{2t} are jointly log-normally distributed, and there is an analytical form available for the expectation of the maximum of two correlated log-normal random variables. In particular, it can be shown that if $\log u_1 \sim N(\mu_1, \sigma_1^2)$, $\log u_2 \sim N(\mu_2, \sigma_2^2)$ and the correlation of $\log u_{1t}$ and $\log u_{2t}$ is ρ_{12}, then the expectation of the maximum of u_{1t} and u_{2t} takes the form:

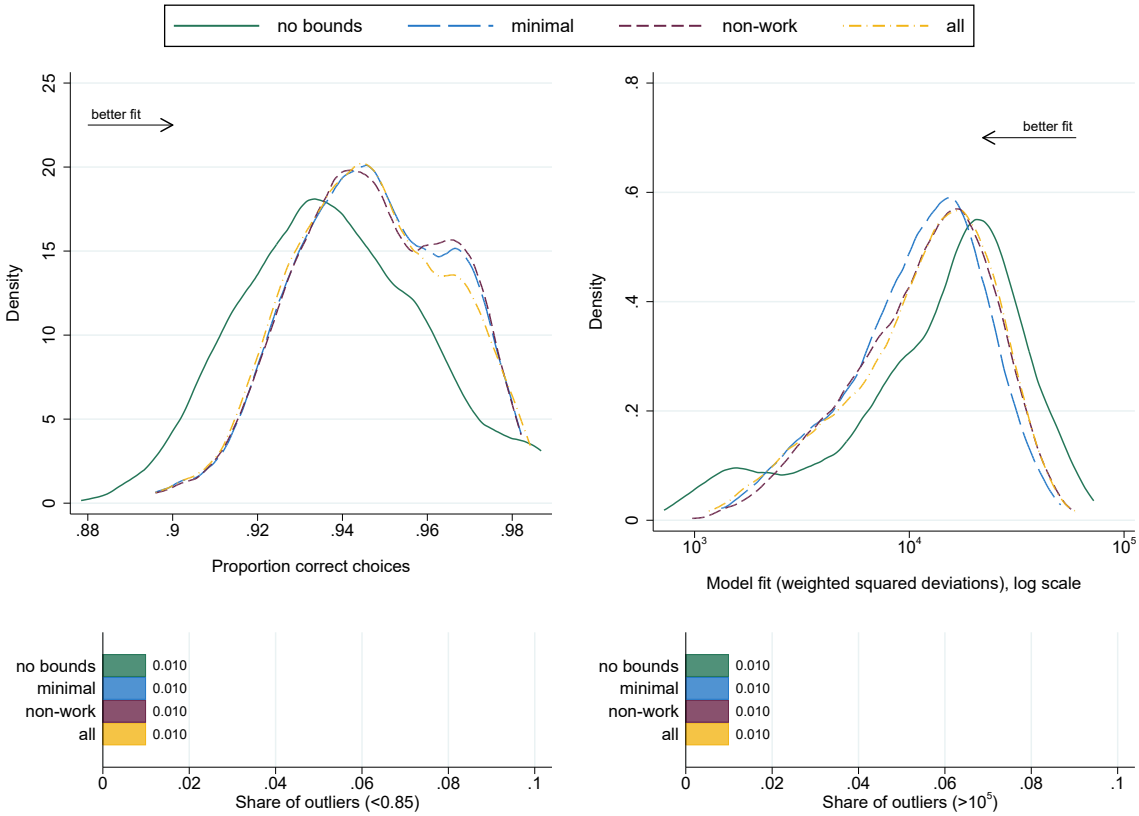
$$\begin{aligned}\mathrm{E}\left[\max(u_{1t}, u_{2t})\right] = \exp\left(\mu_1 + \frac{\sigma_1^2}{2}\right)\Phi\left(\frac{\mu_1 - \mu_2 + \sigma_1^2 - \rho_{12}\sigma_1\sigma_2}{\sqrt{\sigma_1^2 - 2\rho_{12}\sigma_1\sigma_2 + \sigma_2^2}}\right) \\ + \exp\left(\mu_2 + \frac{\sigma_2^2}{2}\right)\Phi\left(\frac{\mu_2 - \mu_1 + \sigma_2^2 - \rho_{12}\sigma_1\sigma_2}{\sqrt{\sigma_1^2 - 2\rho_{12}\sigma_1\sigma_2 + \sigma_2^2}}\right).\end{aligned} \tag{B.4}$$

Figure 10 shows what happens to the simulation accuracy of the traditional KW94 method when different bounds are imposed. The column labelled "no bounds" shows the fit when no bounds are imposed at all. "Minimal" shows the effect of imposing the bound that

$$\underbrace{\mathrm{E}[\max_{k \leq K} V_{kt}(\boldsymbol{S}_t)]}_{\text{EMAX}} \geq \underbrace{\max_{k \leq K} \mathrm{E}\left[V_{kt}(\boldsymbol{S}_t)\right]}_{\text{MAXE}},$$

which follows from a straightforward application of Jensen's Inequality. "Non-work" additionally imposes the bound based on the analytical expression for $\mathrm{E}\left[\max(V_3, V_4)\right]$. "All" also imposes the additional bound for $\mathrm{E}\left[\max(V_1, V_2)\right]$.

Figure 10: Performance of the KW94 method with different bounds



Notes. The densities shown are kernel densities estimated using an Epanechnikov kernel with the bandwidth selected using Silverman's Rule. In each case, the kernel density is calculated over 500 approximation runs with different shocks and approximation draws. In all cases, a maximum of $M = 500$ nodes were evaluated each period, $D = 2000$ draws were used for Monte Carlo integration, and $N = 10,000$ individuals' choices were simulated. "No bounds" shows the fit when no bounds are imposed at all. "Minimal" shows the effect of imposing the bounds that $EMAX \geq MAXE$ that follows from a straightforward application of Jensen's Inequality. "Non-work" additionally imposes the bound based on the analytical expression for $\mathrm{E}\left[\max(V_3, V_4)\right]$. "All" also imposes the additional bound for $\mathrm{E}\left[\max(V_1, V_2)\right]$. The two panels show our two different metrics of approximation performance. Realized shocks $(\boldsymbol{\epsilon}_{it})_{t=1}^{T}$ are held constant only for the evaluation of the proportion of correct choices (left panel).

On the whole, the improvements in fit to be gained from the additional bounds in the canonical model appear to be minor. Only the "minimal" bound that $EMAX \geq MAXE$ leads to a substantial improvement in approximation performance. On the whole, the effect of imposing bounds is not unambiguously positive. One reason is likely to be that what counts for approximation performance is not how closely the approximation reproduces the true EMAX in absolute terms, but only the differences between the nodes. Hence an increase in the approximated value of some nodes but not others as a result of imposing a tighter lower bound can be counterproductive. In sum, it may be best for applied researchers to stick to the "minimal" approach to bounding recommended in KW94.

C Alternative approximation methods

Alternatives to the KW94 approximation method exist and have been used in empirical work. They can be broadly divided into two groups: those that do and do not rely on the MAXE approximation. We discuss each class in turn.[37]

C.1 Methods using MAXE

These methods are built around the observation that in many models, $\max_{k \leq K} \mathrm{E}\left[V_{kt}(\boldsymbol{S}_t)\right]$ ('MAXE') provides a decent approximation to the actual expected value of each state space point $\mathrm{E}\left[\max_{k \leq K} V_{kt}(\boldsymbol{S}_t)\right]$ ('EMAX'). At the same time, MAXE is usually very cheap to calculate, as no numerical integration is required. As a result, these methods can allow for very fast approximation even when the state space is orders of magnitude larger than in the canonical KW94 model.

The most common approach is to simply replace EMAX by MAXE in the computation of continuation values (e.g. Stock and Wise, 1990). This works well in some contexts, but fails in the canonical model: on average, not even a third of choices accord with the 'true' simulated choices. The reason is that differences between EMAX and MAXE are a key driver of decisions in that model, so approximating EMAX by MAXE leads to very different simulated choices.

A potential improvement on this approach resembles the KW94 method. Under this procedure, one approximates the option value of state space points by introducing an auxiliary polynomial into the model representing the difference between MAXE and EMAX. In particular, in the canonical model, we replace each $\hat{\mathrm{E}}\left[V_t(\boldsymbol{S}_t)|\overline{\boldsymbol{S}}_t, d_{kt}=1\right]$ by

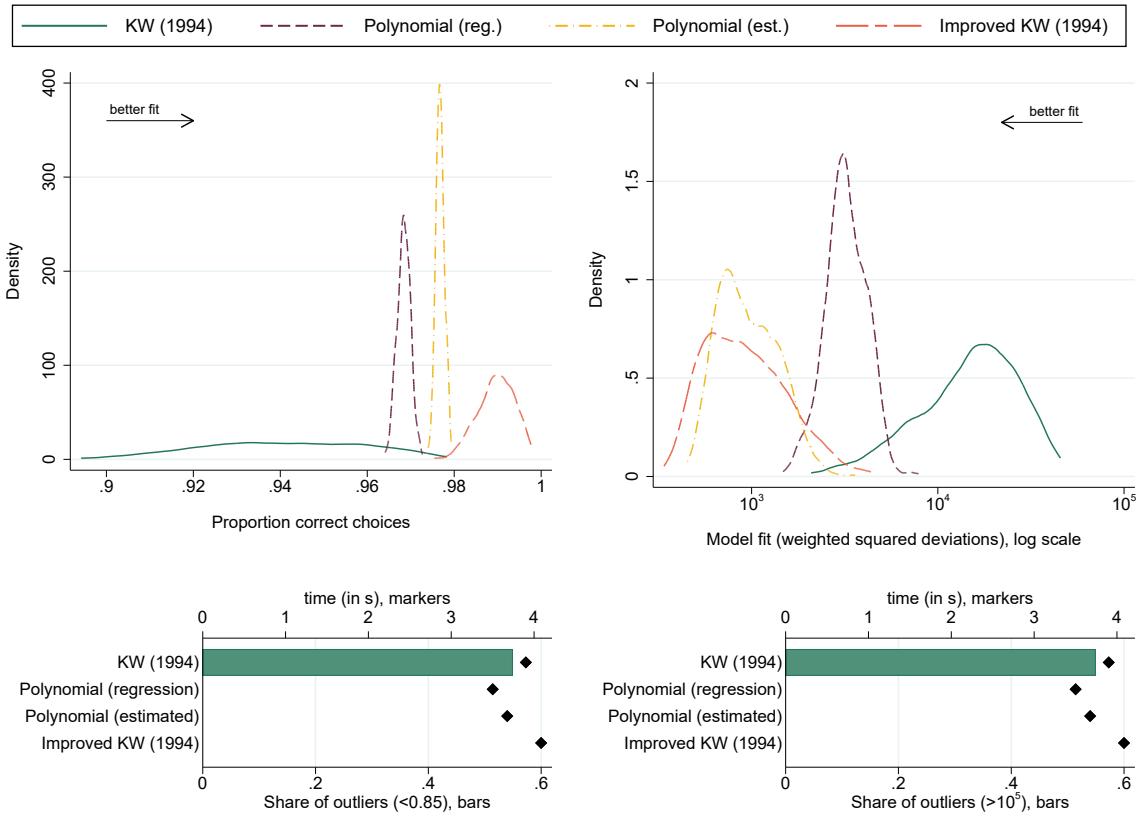
$$\begin{aligned}\check{\mathrm{E}}\left[V_t(\boldsymbol{S}_t)|\overline{\boldsymbol{S}}_t, d_{kt}=1\right] &= \max_{k \leq K} \mathrm{E}\left[V_{kt}(\overline{\boldsymbol{S}}_t, \boldsymbol{\epsilon}_t)\right] \\ +1(s_t < s_{max}) \left(\sum_{j=1}^{K} \pi_j \left[\max_{k \leq K} \mathrm{E}\left[V_{kt}(\boldsymbol{S}_t)\right] - \mathrm{E}\left[V_{jt}(\boldsymbol{S}_t)\right]\right] + \sum_{j=1}^{K} \pi_{K+j} \sqrt{\max_{k \leq K} \mathrm{E}\left[V_{kt}(\boldsymbol{S}_t)\right] - \mathrm{E}\left[V_{jt}(\boldsymbol{S}_t)\right]}\right) \\ + \pi_{2K+1} 1(s_t = s_{max}). \quad \text{(C.1)}\end{aligned}$$

The $2K+1$ elements of $\boldsymbol{\pi}$ are then estimated along with the vector of structural parameters $\boldsymbol{\theta}$.[38]

[37] We do not discuss the performance of the Conditional Choice Probability (CCP) approach (Hotz and Miller, 1993; Arcidiacono and Miller, 2011), as this approach generally requires errors in the model to be additive (although see Kristensen, Nesheim and de Paula, 2015, for arguments why this condition can in principle be relaxed). This is not the case in either of the models discussed in this paper, as the wage functions take a semi-log form.

[38] Note that we do not include an intercept term or functions of the time period t. As the estimation of $\boldsymbol{\pi}$ is solely based on observed choices, these 'scale' parameters are — depending on how they are included — either not

Figure 11: Alternative methods using MAXE: canonical model



Notes. The densities shown are kernel densities estimated using an Epanechnikov kernel with the bandwidth selected using Silverman's Rule. In each case, the kernel density is calculated over 500 approximation runs with different shocks and approximation draws. In all cases, $D = 2000$ draws were used for Monte Carlo integration, and $N = 10,000$ individuals' choices were simulated. "Polynomial (reg.)" shows the performance of the alternative approach if the auxiliary parameter vector π is estimated using OLS regression, using the 'true' values of $\mathrm{E}\left[\max_{k \leq K} V_{kt}(\boldsymbol{S}_t)\right]$ ('EMAX') as the dependent variable and including one observation for each person and period in the simulated data. "Polynomial (est.)" shows the fit when $\boldsymbol{\pi}$ is estimated using the Simulated Method of Moments with the model fit as the criterion function, holding constant the structural parameters $\boldsymbol{\theta}$ at the true values. "KW (1994)" and "Improved KW (1994)" show the performance of the traditional and the improved KW94 methods with $M = 250$ nodes evaluated by Monte Carlo integration for comparison. The two panels show our two different metrics of approximation performance. Realized shocks $(\boldsymbol{\epsilon}_{it})_{t=1}^{T}$ are held constant only for the evaluation of the proportion of correct choices (left panel).

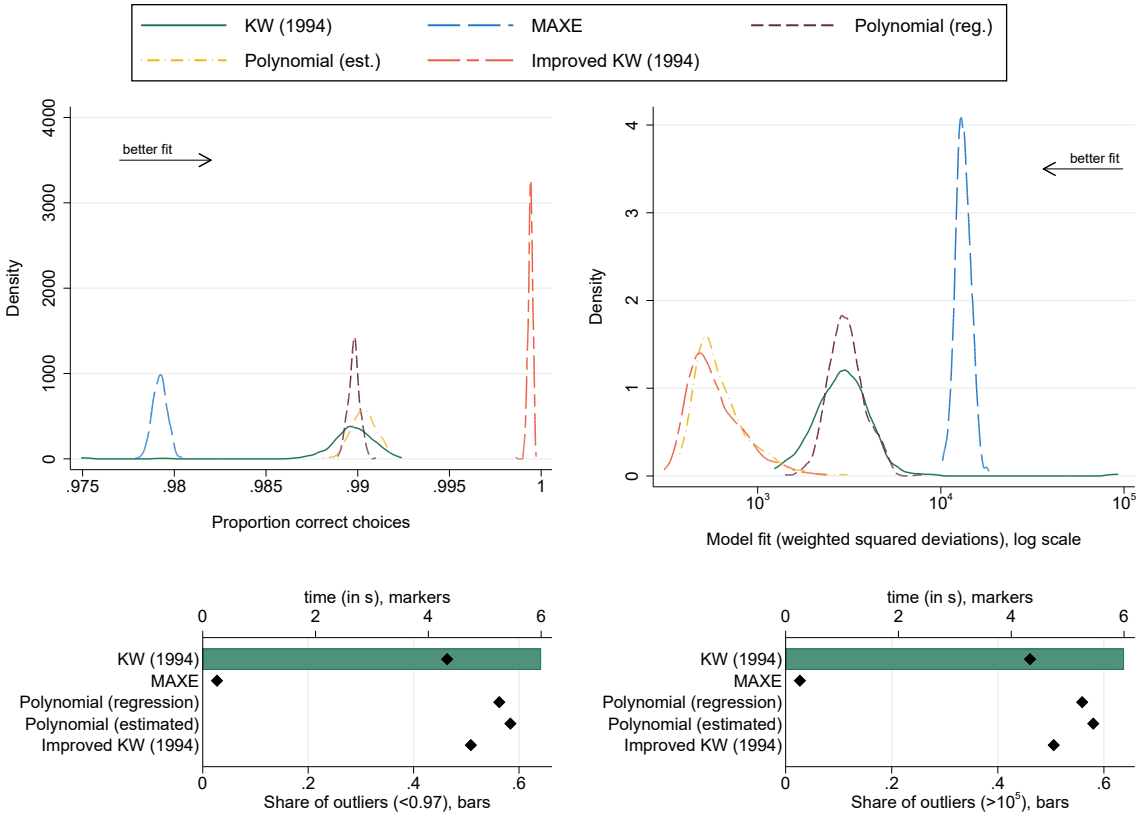
Figure 11 shows the approximation performance of this method compared to both the traditional version and our improved version of the KW94 method (with a maximum of $M = 250$ nodes evaluated to roughly match the other method in terms of computation time). "Polynomial (regression)" shows the performance of the alternative approach if the auxiliary parameter vector $\boldsymbol{\pi}$ is estimated using OLS regression, using the 'true' values of $\mathrm{E}\left[\max_{k \leq K} V_{kt}(\boldsymbol{S}_t)\right]$ ('EMAX') as the dependent variable and including one observation for each person and period in the simulated data.[39] "Polynomial (estimated)" shows the fit when $\boldsymbol{\pi}$ is estimated using the

identified at all or only very weakly identified as a result of the maximum schooling limit.

[39] As scale is not identified (see fn. 38 above), we also include a full set of period dummies in that regression, but

Simulated Method of Moments with the model fit as the criterion function, holding constant the structural parameters $\boldsymbol{\theta}$ at the true values.

Figure 12: Alternative methods using MAXE: model with heterogeneity



Notes. The densities shown are kernel densities estimated using an Epanechnikov kernel with the bandwidth selected using Silverman's Rule. In each case, the kernel density is calculated over 500 approximation runs with different shocks and approximation draws. In all cases, $D = 2000$ draws were used for Monte Carlo integration, and $N = 10,000$ individuals' choices were simulated. "Polynomial (reg.)" shows the performance of the alternative approach if the auxiliary parameter vector $\boldsymbol{\pi}$ is estimated using OLS regression, using the 'true' values of $\mathrm{E}\left[\max_{k \leq K} V_{kt}(\boldsymbol{S}_t)\right]$ ('EMAX') as the dependent variable and including one observation for each person and period in the simulated data. "Polynomial (est.)" shows the fit when $\boldsymbol{\pi}$ is estimated using the Simulated Method of Moments with the model fit as the criterion function, holding constant the structural parameters $\boldsymbol{\theta}$ at the true values. "KW (1994)" and "Improved KW (1994)" show the performance of the traditional and the improved KW94 methods with $M = 250$ nodes evaluated by Monte Carlo integration for comparison. The two panels show our two different metrics of approximation performance. Realized shocks $(\boldsymbol{\epsilon}_{it})_{t=1}^{T}$ are held constant only for the evaluation of the proportion of correct choices (left panel).

Overall, the alternative approach is no match for the improved KW94 method, even when the parameters of the auxiliary parameters are estimated from the 'true' EMAX values, or holding constant the structural parameters at their true values (either of which would be impossible in practice). However, under these idealized conditions, the alternative approach does offer better approximation performance at similar computational cost than the traditional KW94

disregard the estimated period effects.

method with at most $M = 250$ evaluation points. In part, this is because the KW94 method performs poorly in over half of all cases when only $M = 250$ nodes are evaluated using Monte Carlo integration.

Figure 12 is the equivalent figure for the model with heterogeneity. We include the approximation of EMAX by MAXE in this figure, as in the model with heterogeneity, that approximation also offers acceptable performance. All methods perform much better, as more *ex ante* heterogeneity means that the precise continuation value of each state space point matters less in shaping agents' decisions.

The improved KW94 method still dominates in terms of approximation performance with a virtually perfect fit. The method with an auxiliary polynomial again yields better approximation performance than the traditional KW94 method. The simple MAXE method leads to a somewhat worse fit, but is much faster to compute. This suggests that in models with extremely large state spaces and substantial heterogeneity, it might be advisable to estimate the model using the simple MAXE approximation in an initial step.[40]

C.2 Methods not using MAXE

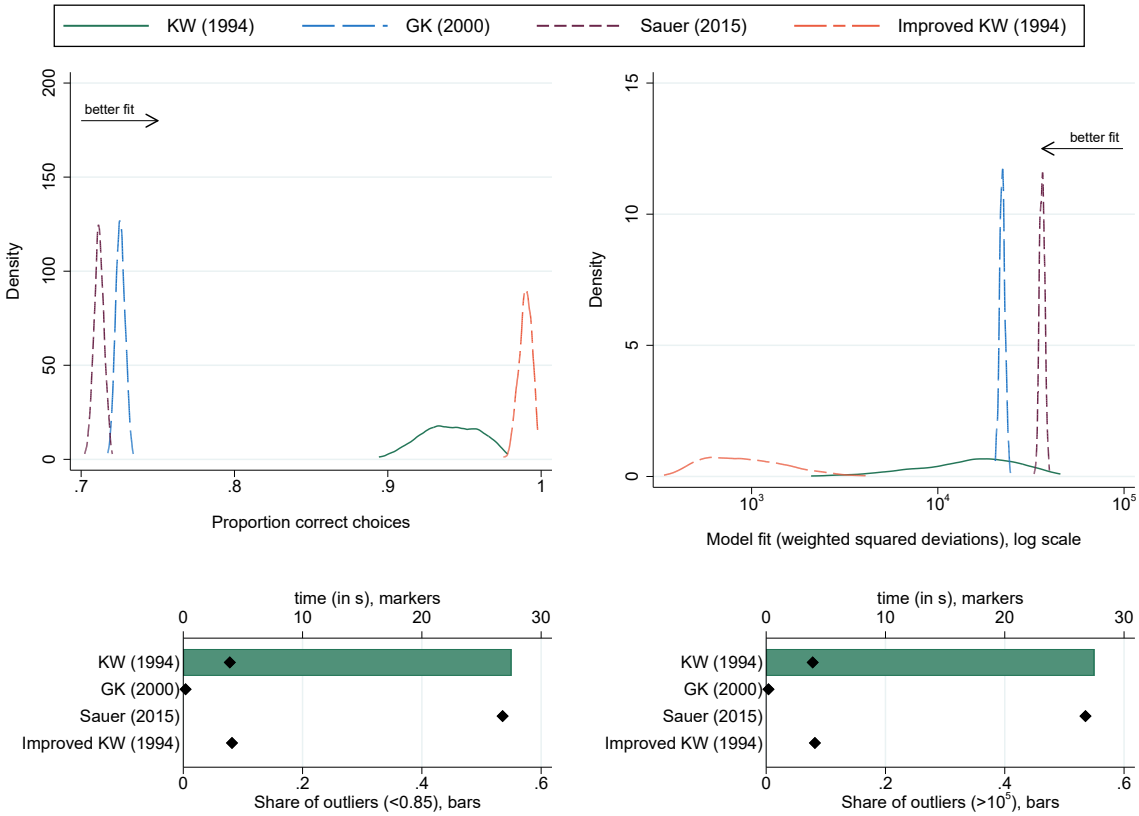
Other approximation methods for $\mathrm{E}\left[\max_{k \leq K} V_{kt}(\boldsymbol{S}_t)\right]$ ('EMAX') do not make use of the MAXE approximation at all. The simplest method is to just replace all continuation values by a polynomial in the pre-determined state variables, and estimate the parameters of that polynomial along with the structural parameters of the model. An approximation of this kind is proposed in Geweke and Keane (2000).

A variation on this approach is the method of Sauer (2015) (also used in Belzil, Hansen and Liu, 2017). Instead of approximating the continuation value $EMAX_{t+1}$ by polynomial approximation, this method uses Monte Carlo integration to evaluate $EMAX_{t+1}$, but instead of proceeding by backward recursion, the subsequent period's continuation value $EMAX_{t+2}$ is approximated by a polynomial in the pre-determined state variables. This means that unlike for the full solution method, it is not necessary to evaluate all possible state space nodes. Only the at most $KN(T-2)$ nodes that are directly reachable from the nodes actually reached by individuals in the data need to be evaluated by Monte Carlo Integration.[41]

[40] Whether the method with an auxiliary polynomial has a useful role to play in some contexts is an interesting question for further research.

[41] This is an upper bound, as duplicate nodes only need to be evaluated once. As a result, computation time can vary substantially with both structural and auxiliary parameters.

Figure 13: Alternative methods without MAXE: canonical model



Notes. The densities shown are kernel densities estimated using an Epanechnikov kernel with the bandwidth selected using Silverman's Rule. In each case, the kernel density is calculated over 500 approximation runs with different shocks and approximation draws. In all cases, $D = 2000$ draws were used for Monte Carlo integration, and $N = 10,000$ individuals' choices were simulated. "GK (2000)" and "Sauer (2015)" relate to the Geweke and Keane (2000) and Sauer (2015) methods, respectively. In each case, the vector of auxiliary parameters $\boldsymbol{\pi}$ was estimated using the Simulated Method of Moments with the model fit as the criterion function, holding constant the structural parameters $\boldsymbol{\theta}$ at the true values. "KW (1994)" and "Improved KW (1994)" show the performance of the traditional and the improved KW94 methods with $M = 250$ nodes evaluated by Monte Carlo integration for comparison. The two panels show our two different metrics of approximation performance. Realized shocks $(\boldsymbol{\epsilon}_{it})_{t=1}^{T}$ are held constant only for the evaluation of the proportion of correct choices (left panel).

Figure 13 shows results from the canonical model for both the Geweke and Keane (2000) method (labelled "GK (2000)") and the Sauer (2015) method (labelled "Sauer (2015)").[42] In general, these appear to provide a much less accurate approximation than either the KW94 method or the methods using MAXE described in the previous section, although this may be

[42]The polynomials used for both methods are

$$\tilde{\tilde{EMAX}}_t = \pi_1 x_{1t} + \pi_2 x_{1t}^2 + \pi_3 x_{2t} + \pi_4 x_{2t}^2 + \pi_5 s_t + \pi_6 s_t^2 + \pi_7 d_{3,t-1}.$$

for the canonical KW94 model and

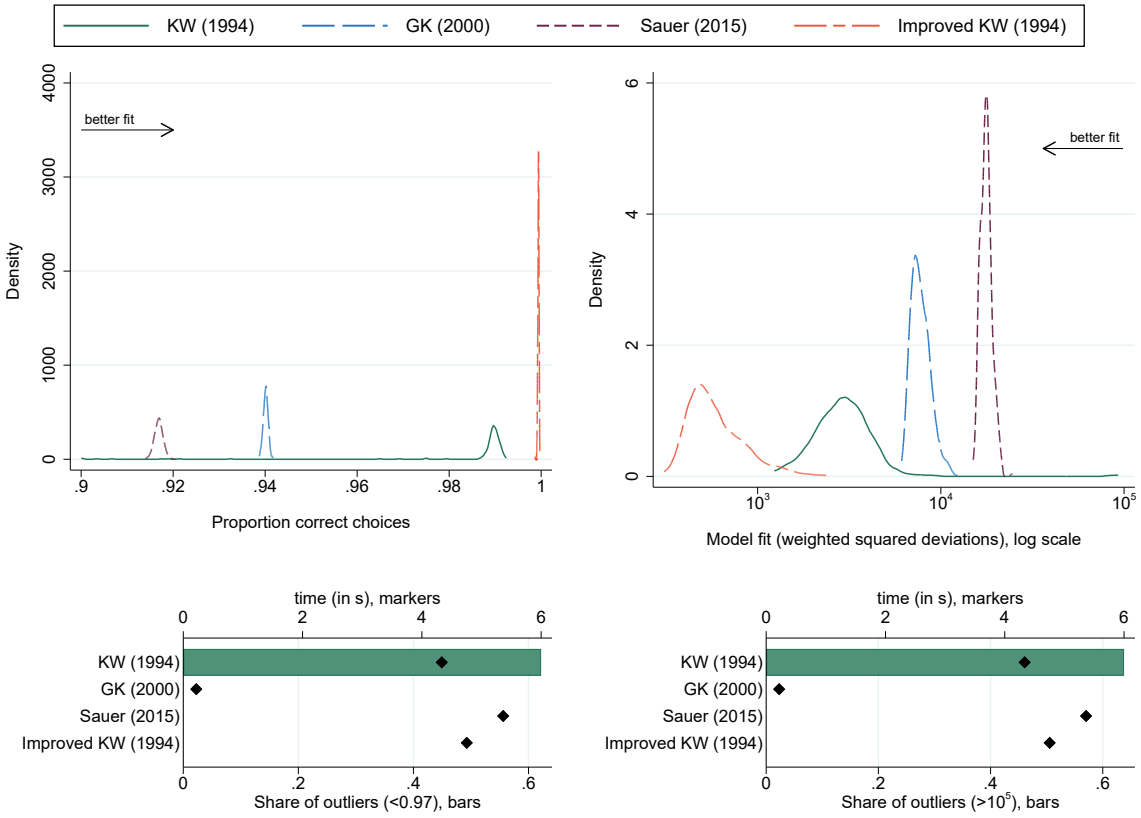
$$\tilde{\tilde{EMAX}}_t = \pi_1 x_{1t} + \pi_2 x_{1t}^2 + \pi_3 x_{2t} + \pi_4 x_{2t}^2 + \pi_5 s_t + \pi_6 s_t^2.$$

for the model with heterogeneity. $d_{3,t-1}$ was excluded for the model with heterogeneity because the parameter π_7 was poorly identified. No intercept or period effects were included, as they cannot in general be identified (see fn. 38).

a consequence of the particular polynomials selected.[43]

Interestingly, the approximation fit is somewhat better using the Geweke and Keane (2000) method than using the Sauer (2015) method, which suggests that the computationally costly inclusion of an additional period of state space points evaluated by Monte Carlo integration can *worsen* the approximation fit. However, it should be noted that this need not lead to worse estimation performance, as the additional period of state space points evaluated by Monte Carlo integration also provides an additional source of identification.

Figure 14: Alternative methods without MAXE: model with heterogeneity

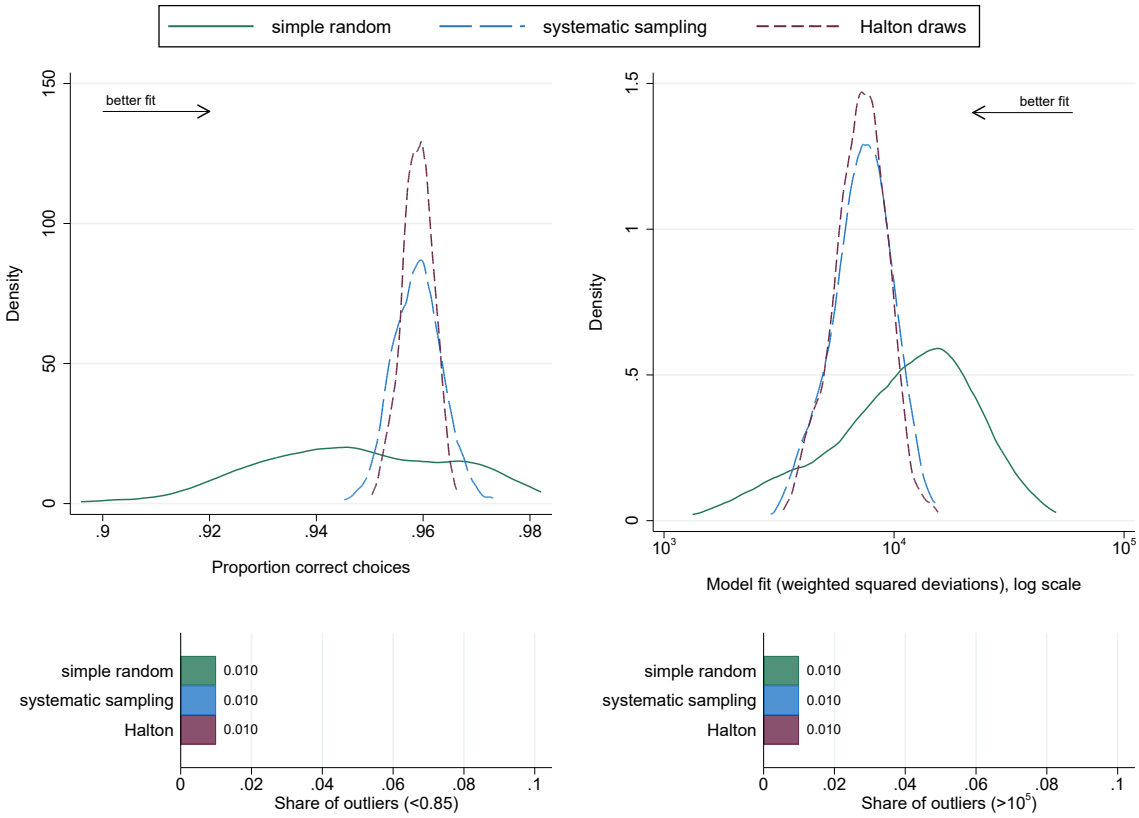


Notes. The densities shown are kernel densities estimated using an Epanechnikov kernel with the bandwidth selected using Silverman's Rule. In each case, the kernel density is calculated over 500 approximation runs with different shocks and approximation draws. In all cases, $D = 2000$ draws were used for Monte Carlo integration, and $N = 10,000$ individuals' choices were simulated. "GK (2000)" and "Sauer (2015)" relate to the Geweke and Keane (2000) and Sauer (2015) methods, respectively. In each case, the vector of auxiliary parameters $\boldsymbol{\pi}$ was estimated using the Simulated Method of Moments with the model fit as the criterion function, holding constant the structural parameters $\boldsymbol{\theta}$ at the true values. "KW (1994)" and "Improved KW (1994)" show the performance of the traditional and the improved KW94 methods with $M = 250$ nodes evaluated by Monte Carlo integration for comparison. The two panels show our two different metrics of approximation performance. Realized shocks $(\boldsymbol{\epsilon}_{it})_{t=1}^{T}$ are held constant only for the evaluation of the proportion of correct choices (left panel).

[43]There appears to be a trade-off between approximation accuracy (achieved by including a polynomial with a large number of auxiliary parameters) and the identification of the structural parameters, which will likely hinge on the number of auxiliary parameters being relatively small. The polynomial selected here is roughly of the same complexity as the one used in Sauer (2015).

Figure 14 provides the analogous results for the model with *ex ante* heterogeneity. Again approximation performance is worse than using the KW94 method or alternative methods using MAXE. These results suggest that applied researchers will need to be careful before applying these methods, as with standard estimation methods and a relatively simple polynomial, the approximation performance appears insufficient to guarantee the reliable estimation of complex models.

Figure 15: Performance of the KW94 method with different sampling methods



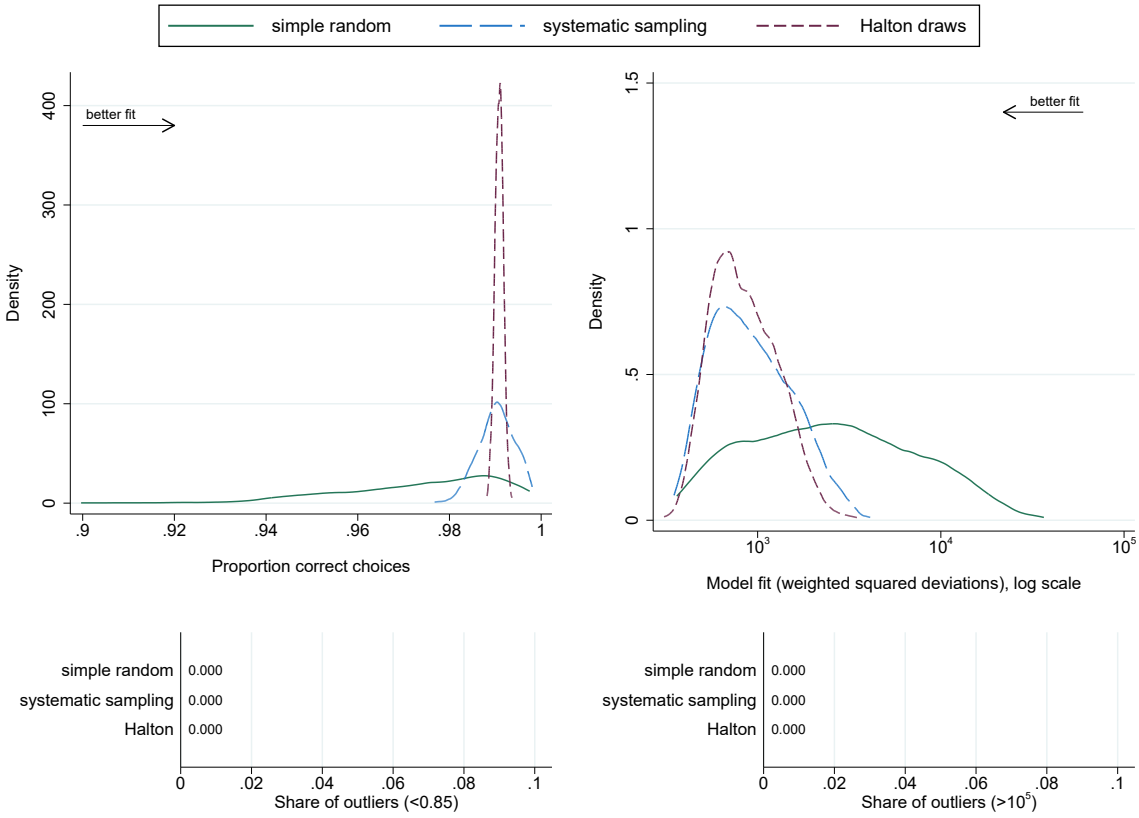
Notes. The densities shown are kernel densities estimated using an Epanechnikov kernel with the bandwidth selected using Silverman's Rule. In each case, the kernel density is calculated over 500 approximation runs with different shocks and approximation draws. In all cases, a maximum of $M = 500$ nodes were evaluated each period, $D = 2000$ draws were used for Monte Carlo integration, and $N = 10,000$ individuals' choices were simulated. "Simple" refers to simple random sampling, "Systematic" refers to the systematic sampling algorithm outlined in section 3.1, and Halton draws are obtained following the algorithm from Halton (1964). The two panels show our two different metrics of approximation performance. Realized shocks $(\boldsymbol{\epsilon}_{it})_{t=1}^{T}$ are held constant only for the evaluation of the proportion of correct choices (left panel).

D Comparison of different sampling methods

D.1 KW94 method

Figure 15 shows the influence of different sampling methods on the performance of the traditional version of the KW94 method in the canonical model with a maximum of $M = 500$ evaluation points. Simple random draws are compared to the systematic sampling method outlined in section 3.1 and Halton (1964) systematic draws. Both systematic sampling and Halton draws lead to large improvements in approximation accuracy. Halton draws are slightly better than systematic sampling, but the differences are minor.[44]

Figure 16: Performance of the full solution method with different sampling methods



Notes. The densities shown are kernel densities estimated using an Epanechnikov kernel with the bandwidth selected using Silverman's Rule. In each case, the kernel density is calculated over 500 approximation runs with different shocks and approximation draws. In all cases, $D = 2000$ draws were used for Monte Carlo integration, and $N = 10,000$ individuals' choices were simulated. "Simple" refers to simple random sampling, "Systematic" refers to the systematic sampling algorithm outlined in section 3.1, and Halton draws are obtained following the algorithm from Halton (1964). The two panels show our two different metrics of approximation performance. Realized shocks $(\boldsymbol{\epsilon}_{it})_{t=1}^{T}$ are held constant only for the evaluation of the proportion of correct choices (left panel).

[44] Intuitively, the reason that Halton draws perform somewhat better is that they ensure even integral coverage in multiple dimensions, whereas our systematic sampling method only ensures even coverage in each dimension individually.

D.2 Full solution

Figure 16 shows the influence of different sampling methods on the performance of the full solution method in the canonical KW94 model. Accuracy gains are even larger than when using the KW94 method, with both systematic sampling and Halton draws offering an excellent fit. Using the full solution method, the advantage of Halton draws is somewhat larger in relative terms.[45]

Table 4: Average deviation (bias) of estimated parameters in the canonical model

		Average deviation from true parameter				
Parameter	**True value**	**Full solution**	**KW250**	**KW250 new**	**KW500**	**KW500 new**
α_{00}	9.21	.00464	-.00273	.00409	.00512	.00494
α_{01}	.038	-.00012	.00071	-9.5e-05	-9.4e-05	-.00014
α_{02}	.033	-.00034	-.00057	-.0003	-.00039	-.00035
α_{03}	.0005	-9.1e-06	-1.7e-05	-8.4e-06	-1.0e-05	-9.2e-06
α_{04}	0	-.00019	-.0001	-.00017	-.00049	-.00022
α_{05}	0	-5.1e-06	5.0e-05	-3.4e-06	-1.4e-05	-6.1e-06
α_{10}	8.48	.00392	.01703	.0042	.0056	.0042
α_{11}	.07	-.00016	-.00119	-.00021	-.00035	-.0002
α_{12}	.067	-7.3e-05	.00139	4.2e-05	.00011	1.5e-05
α_{13}	.001	8.4e-08	3.7e-05	3.8e-06	4.8e-06	3.0e-06
α_{14}	.022	-.00018	-.00338	-.00025	-.00045	-.00022
α_{15}	.0005	2.1e-06	-.00013	-1.1e-06	-1.5e-05	5.7e-08
β_0	0	77.502	-39.536	84.26	-3.8551	90.82
β_1	0	31.616	101.55	34.448	27.902	37.208
β_2	4000	-131.72	-1500.5	-137.01	-159.78	-131.77
γ_{0_f}	17750	68.271	54.957	67.283	54.334	67.768
a_{11}	.2	-.00107	-.00098	-.00091	-.00099	-.00104
a_{22}	.25	1.2e-05	-.0005	2.8e-05	-9.0e-05	2.7e-05
a_{33}	1500	-372.81	574.77	-371.67	-378.99	-372.13
a_{44}	1500	-246.99	613.65	-252.8	-249.09	-237.67

Notes. The average deviations from the true parameters were calculated over 24 estimation runs as described in the text, using simulated balanced panels of $N = 2000$ individuals each. $D = 2000$ solution draws were used in model solution. For each observation, 200 Halton draws were used to simulate the likelihood. Likelihood simulation draws were held constant across estimation runs in order to minimize statistical noise unrelated to the different approximation methods.

E Bias and standard deviation of estimated parameters

E.1 Canonical model: Simulated Maximum Likelihood

Table 4 shows the average deviation of the estimated parameters from the true parameters. For the full solution, these average deviations are mostly quite small relative to the scale of each parameter. The exception are the standard deviations of the non-work options a_{33} and a_{44}, which are substantially underestimated. This is a known artefact of Simulated Maximum

[45] Intuitively, even integral coverage becomes more important when other sources of approximation error have been eliminated.

Likelihood estimation in this model; the same bias is found in KW94.[46] Biases for the improved versions of the KW94 method are generally very close to the full solution. They are mostly larger for the traditional KW94 method, especially when only a maximum of $M = 250$ state space points are evaluated by Monte Carlo integration.

Table 5 show the sample standard deviation of the estimated parameters. Sample standard deviations tend to be of similar orders of magnitude as the average deviation from the true parameters. Standard deviation methods for all approximation methods are similar to the full solution, except for the traditional KW94 method with $M = 250$, which has much larger standard deviations. This appears to be driven by a minority of estimation runs that do not converge to parameter values near the true parameters (not shown).

Table 5: Sample standard deviation of estimated parameters in the canonical model

		Sample standard deviation				
Parameter	**True value**	**Full solution**	**KW250**	**KW250 new**	**KW500**	**KW500 new**
α_{00}	9.21	.0081	.01564	.00837	.00869	.00846
α_{01}	.038	.00071	.0016	.00075	.00075	.00073
α_{02}	.033	.00032	.00069	.0003	.00033	.00032
α_{03}	.0005	1.0e-05	2.0e-05	9.8e-06	1.1e-05	1.0e-05
α_{04}	0	.00034	.00112	.00033	.00028	.00034
α_{05}	0	1.7e-05	8.6e-05	1.7e-05	1.6e-05	1.7e-05
α_{10}	8.48	.0041	.01904	.00386	.00428	.00416
α_{11}	.07	.00028	.00143	.00029	.00028	.00027
α_{12}	.067	.00035	.00172	.00037	.00037	.00035
α_{13}	.001	1.1e-05	4.4e-05	1.2e-05	1.2e-05	1.1e-05
α_{14}	.022	.00014	.0038	.00016	.00014	.00014
α_{15}	.0005	1.2e-05	.00015	1.3e-05	1.2e-05	1.2e-05
β_0	0	132.63	221.73	133.55	142.38	128.2
β_1	0	77.847	279.86	80.672	67.949	78.504
β_2	4000	57.199	1713.6	54.389	58.261	55.63
γ_{0_f}	17750	67.211	68.543	64.427	69.477	68.538
a_{11}	.2	.00121	.00126	.00118	.00123	.00121
a_{22}	.25	.00075	.00163	.00076	.00076	.00076
a_{33}	1500	56.649	1369.9	54.182	55.83	55.731
a_{44}	1500	94.477	1209.9	96.81	100.84	94.769

Notes. The sample standard deviations were calculated over 24 estimation runs as described in the text, using simulated balanced panels of $N = 2000$ individuals each. $D = 2000$ solution draws were used in model solution. For each observation, 200 Halton draws were used to simulate the likelihood. Likelihood simulation draws were held constant across estimation runs in order to minimize statistical noise unrelated to the different approximation methods.

E.2 Canonical model: Simulated Method of Moments

Tables 6 and 7 are the same tables for the Simulated method of Moments (SMM) estimation. Both average deviations from the truth and sample standard deviations are somewhat larger

[46] This is likely related to the use of the smoothed-Logit simulator (McFadden, 1989), which is known to introduce bias (cf. the discussion in Keane and Wolpin, 1994). In particular, the smoothed-Logit simulator tends to overstate the likelihood of tail events, which may explain a downward bias in the error variance.

for SMM, likely reflecting lower efficiency compared to Simulated Maximum Likelihood. The relative performance of different approximation methods is similar to the maximum likelihood results, with the improved KW94 method performing both better and similar to the full solution.

Table 6: Average deviation (bias) of estimated parameters in the canonical model

		Average deviation from true parameter				
Parameter	**True value**	**Full solution**	**KW250**	**KW250 new**	**KW500**	**KW500 new**
α_{00}	9.21	-.00401	-.00487	-.00428	-.00337	-.00404
α_{01}	.038	.00041	.00246	.00037	.0005	.0004
α_{02}	.033	-8.1e-05	-.00136	-2.2e-06	-1.6e-05	-7.8e-05
α_{03}	.0005	-2.1e-06	-2.8e-05	-3.1e-07	-8.1e-07	-3.2e-06
α_{04}	0	-2.4e-05	-.00204	-.00025	-.00016	.00013
α_{05}	0	1.0e-06	-6.8e-05	-1.2e-05	3.0e-06	7.8e-06
α_{10}	8.48	-.00284	.02545	-.00204	.00478	-.00326
α_{11}	.07	.00044	-.00053	.00034	-.00029	.00048
α_{12}	.067	-.00027	-.00119	-.00018	.00011	-.00023
α_{13}	.001	-5.8e-06	-2.3e-05	-3.5e-06	4.1e-06	-5.2e-06
α_{14}	.022	9.7e-05	-.00131	-7.6e-06	-.0002	.0001
α_{15}	.0005	3.2e-05	-4.7e-05	1.8e-05	8.2e-06	3.5e-05
β_0	0	-69.664	-2378.2	-71.489	10.043	-77.732
β_1	0	135.36	-1598.7	81.936	50.341	144.48
β_2	4000	-136.09	3561	-156.03	-156.2	-150.1
γ_{0_f}	17750	129.96	-1923.7	171.41	-113.5	132.29
a_{11}	.2	.00024	-.00623	.00065	-.00149	3.2e-05
a_{22}	.25	-.00133	-.00335	-.0011	-.00055	-.00157
a_{33}	1500	-493.64	2968.3	-548.88	-490.34	-528
a_{44}	1500	-645.01	1019.1	-751.39	-143.13	-666.45

Notes. The average deviations from the true parameters were calculated over 24 estimation runs as described in the text, using simulated balanced panels of $N = 10,000$ individuals each. $D = 2000$ solution draws were used in model solution.

Table 7: Sample standard deviation of estimated parameters in the canonical model

		Sample standard deviation				
Parameter	**True value**	**Full solution**	**KW250**	**KW250 new**	**KW500**	**KW500 new**
α_{00}	9.21	.01359	.02088	.01574	.01479	.01782
α_{01}	.038	.00128	.00301	.00142	.00136	.00148
α_{02}	.033	.00042	.00198	.00032	.00032	.00026
α_{03}	.0005	1.4e-05	4.4e-05	9.8e-06	9.5e-06	8.2e-06
α_{04}	0	.00118	.00366	.00074	.00111	.00106
α_{05}	0	6.5e-05	.00014	4.4e-05	5.5e-05	5.4e-05
α_{10}	8.48	.00543	.03507	.00715	.00789	.00631
α_{11}	.07	.00052	.00297	.0007	.00084	.00062
α_{12}	.067	.00033	.00347	.00042	.00047	.00032
α_{13}	.001	1.1e-05	8.6e-05	1.2e-05	1.3e-05	9.4e-06
α_{14}	.022	.00059	.00184	.0005	.00042	.00051
α_{15}	.0005	7.6e-05	.0002	5.7e-05	5.1e-05	6.4e-05
β_0	0	240.19	4366.9	268.42	248.55	269.44
β_1	0	193	4180.2	195.47	266.82	163.09
β_2	4000	101.68	6656.6	88.229	84.553	72.841
γ_{0_f}	17750	235.99	4207.4	188.49	340.39	217.86
a_{11}	.2	.00272	.00816	.00209	.00277	.00227
a_{22}	.25	.00126	.00887	.00143	.0015	.00125
a_{33}	1500	136.79	5370.6	155.29	121.98	105.89
a_{44}	1500	585.27	2280.9	556.62	707.27	619.75

Notes. The sample standard deviations were calculated over 24 estimation runs as described in the text, using simulated balanced panels of $N = 10,000$ individuals each. $D = 2000$ solution draws were used in model solution.

E.3 Model with heterogeneity

Tables 8 and 9 are the analogous tables for the model with heterogeneity. Both average deviations from the truth and sample standard deviations are comparable to the canonical model. The heterogeneity parameters ζ_1 and ζ_2 are precisely estimated. The relative performance of different approximation methods is similar to the results for the canonical model, with the improved KW94 method performing both better and similar to the full solution.

Table 8: Average deviation (bias) of estimated parameters in the model with heterogeneity

		Average deviation from true parameter				
Parameter	**True value**	**Full solution**	**KW250**	**KW250 new**	**KW500**	**KW500 new**
α_{00}	9.21	.00268	-.04773	.00205	.00648	-.00083
α_{01}	.038	-.00013	.00339	-6.3e-05	-.00027	-4.1e-05
α_{02}	.033	-2.9e-05	.00156	-9.6e-05	-8.0e-06	-9.9e-05
α_{03}	.0005	-1.0e-06	2.7e-05	-2.7e-06	-1.2e-06	-2.7e-06
α_{04}	0	-.00051	.00064	-.00055	-.00057	-.00056
α_{05}	0	-7.3e-06	.00013	-1.4e-05	2.7e-06	-1.6e-05
α_{10}	8.48	-.00294	.01362	-.00363	.00556	-.00249
α_{11}	.07	-.00016	-.00044	-8.6e-05	-.00079	-.00026
α_{12}	.067	.00033	.0013	.00047	.00065	.00038
α_{13}	.001	1.0e-05	3.6e-05	1.5e-05	1.9e-05	1.3e-05
α_{14}	.022	-.00019	-.00373	-.00024	-.00065	-.00013
α_{15}	.0005	-6.3e-06	-9.7e-05	-8.7e-06	-2.0e-05	-7.4e-06
β_0	0	20.062	-77.889	-82.077	17.9	-85.364
β_1	0	-32.333	-154.04	-87.964	-89.071	-175.31
β_2	4000	-42.963	-674.81	-60.823	-90.592	-45.135
γ_{0_f}	17750	53.26	517	54.197	54.645	49.307
a_{11}	.2	8.8e-05	.00239	.00024	4.7e-05	.00017
a_{22}	.25	.00084	8.2e-05	.00111	.00029	.00073
a_{33}	1500	-363.2	299.92	-355.42	-354.28	-354.49
a_{44}	1500	-431.49	-44.401	-405.29	-431.6	-430.73
ζ_1	.05	-.00158	-.00449	-.00115	-.00467	.00157
ζ_2	.1	.00325	.00216	.00158	.00268	.00422

Notes. The average deviations from the true parameters were calculated over 15 estimation runs as described in the text, using simulated balanced panels of $N = 2000$ individuals each. $D = 2000$ solution draws were used in model solution. For each observation, 200 Halton draws were used to simulate the likelihood. Likelihood simulation draws were held constant across estimation runs in order to minimize statistical noise unrelated to the different approximation methods.

Table 9: Sample standard deviation of estimated parameters in the model with heterogeneity

		Sample standard deviation				
Parameter	**True value**	**Full solution**	**KW250**	**KW250 new**	**KW500**	**KW500 new**
α_{00}	9.21	.01781	.10606	.01772	.03479	.02002
α_{01}	.038	.00112	.00598	.00127	.0023	.00124
α_{02}	.033	.0002	.00186	.00018	.00035	.00025
α_{03}	.0005	4.5e-06	4.1e-05	3.6e-06	8.7e-06	5.6e-06
α_{04}	0	.00042	.00168	.00054	.00064	.0005
α_{05}	0	2.3e-05	.00018	3.7e-05	4.4e-05	3.1e-05
α_{10}	8.48	.00858	.04519	.00768	.0068	.00806
α_{11}	.07	.00043	.00174	.00041	.00067	.00044
α_{12}	.067	.00029	.00127	.00046	.00061	.00036
α_{13}	.001	1.0e-05	3.6e-05	1.5e-05	1.8e-05	1.3e-05
α_{14}	.022	.00023	.00482	.00023	.00039	.00023
α_{15}	.0005	1.6e-05	.00018	1.6e-05	2.0e-05	1.5e-05
β_0	0	338.93	773.99	307.36	358.7	332.58
β_1	0	330.2	720.72	313.57	386.94	319.6
β_2	4000	55.69	910.25	60.284	71.784	56.798
γ_{0_f}	17750	43.798	659.24	54.776	51.562	41.728
a_{11}	.2	.00082	.00777	.00081	.00096	.00087
a_{22}	.25	.00113	.00864	.00087	.0011	.00115
a_{33}	1500	77.369	1034.9	51.414	70.033	81.722
a_{44}	1500	62.034	1077.8	118.4	70.705	60.681
ζ_1	.05	.00808	.06834	.00681	.01552	.00973
ζ_2	.1	.00601	.06966	.00764	.009	.00572

Notes. The sample standard deviations were calculated over 15 estimation runs as described in the text, using simulated balanced panels of $N = 2000$ individuals each. $D = 2000$ solution draws were used in model solution. For each observation, 200 Halton draws were used to simulate the likelihood. Likelihood simulation draws were held constant across estimation runs in order to minimize statistical noise unrelated to the different approximation methods.

Conclusion

This thesis has shown that more efficient estimation and approximation techniques can make a large difference to the performance of different methods for estimating finite-horizon discrete choice dynamic programming models. The techniques introduced in this thesis should enable researchers to estimate more complex models at the same level of accuracy — or estimate a given model at a higher level of accuracy — than has previously been possible. It is hoped that this work will contribute to a broader research agenda seeking greater robustness of structural results in applied microeconomics.

Authorship Declaration

I declare that the first paper of this thesis, entitled "Efficient Likelihood Simulation in Discrete Choice Models with Observed Payoffs", the introduction to this thesis, and the conclusion of this thesis are entirely my own work. (In the interest of full disclosure, I note that some of the MATLAB code used to obtain the results in section 7 of the first paper builds on code originally written by Jack Britton.)

The second paper, entitled "Generalized indirect inference in dynamic structural models: an application to the long-run effect of a conditional cash transfer scheme", is joint work with Jack Britton. I have independently developed the extended Generalized Indirect Inference method discussed in the paper and have independently written the relevant sections of the paper. In addition, I have written and run all of the estimation code for the paper (in the case of section 3 based on initial MATLAB code written by Jack Britton). The applied section (section 4) was developed jointly, based on Jack Britton's earlier draft entitled "Getting poor kids to stay in school: Do conditional cash transfers help?". Some material from that draft has been incorporated into our paper.

The third paper, entitled "Revisiting the solution of dynamic discrete choice models: time to bring back Keane and Wolpin (1994)?" is also joint work with Jack Britton. A version of this paper has recently been published as IFS Working Paper W21/13. I have independently developed all three of the improvements to the Keane and Wolpin (1994) method discussed in the paper, have independently performed all simulation and estimation exercises in the paper, and have written the paper and most of the code underlying it. Jack contributed initial MATLAB code for the approximate solution of the Keane and Wolpin (1994) model as well as some Stata code for the visualization of results, developed our second measure of model fit based on key moments of the observed data, and helped with the framing of the paper.

None of the papers in this thesis are based on my MPhil thesis.

Ben Waltmann

2 June 2021