JOURNAL OF APPLIED ECONOMETRICS
J. Appl. Econ. **15**: 595–624 (2000)

SERIALLY CORRELATED VARIABLES IN DYNAMIC, DISCRETE CHOICE MODELS

TODD R. STINEBRICKNER*

Department of Economics, The University of Western Ontario, Social Science Centre, London, Ontario, Canada N6A 5C2

SUMMARY

This paper discusses the problems that are encountered when dynamic, discrete choice models are specified with continuous, serially correlated state variables. A variety of approximation methods that can deal with these problems is examined, and an empirical example that allows continuous variables to be serially correlated is presented. Copyright © 2000 John Wiley & Sons, Ltd.

1. INTRODUCTION

In many economic contexts, dynamic discrete choice (DDC) models represent a theoretically desirable way to describe an individual's decision making process when the individual faces uncertainty about the future and is forward looking. Unfortunately, the use of certain appealing specifications of such models can often be burdensome or infeasible due to the computational obstacles that are encountered during estimation. The estimation of dynamic, discrete choice models with continuous, serially correlated stochastic components represents one prominent example in which the difficulties which arise during estimation have remained largely unresolved (Rust and Phelan, 1997). One can potentially circumvent these technical difficulties by discretizing the continuous variable(s) of interest and using the stochastic process associated with the continuous variable to generate the appropriate inter-period transition probability matrix for the discretized variable.[1] However, under this approach, measurement error is introduced if the researcher's model treats the discrete variable as if it is the continuous variable of interest, and desirable variation in the data is lost if the discrete variable enters the model as a series of indicator variables. While these problems can be attenuated by discretizing the variable into a larger number of nodes, this will lead to increases in the amount of computer time that is needed for estimation.[2] As a result, it may often be desirable to avoid discretization by treating the variable as continuous.[3]

*Correspondence to: Todd R. Stinebrickner, Department of Economics, University of Western Ontario, Social Science Centre, London, Ontario, Canada N6A 5C2. e-mail: trstineb@julian.uwo.ca

[1] For example, Hubbard *et al.* (1995) allow two state variables to follow first-order autoregressive processes around a deterministic trend by discretizing the deviations from the trend into nine nodes.

[2] The increase in estimation time arises because computing a person's expected future utility involves computing a weighted average of the expected utilities that the person would receive given each of the possible values of the discretized variable. If the model has been specified with a series of indicator variables, increasing the number of nodes also leads to increases in the number of parameters that need to be estimated.

Received 30 September 1999
Revised 10 September 2000

The basic problem that arises when continuous, serially correlated variables are present is that value functions cannot typically be solved exactly. A variety of approximation methods are available to address this problem. However, by and large, these methods were not introduced specifically for the purpose of estimating dynamic, discrete choice models with serially correlated state variables, and only a very limited number of such models have been estimated.[4] As a result, although theoretically these methods can deal with the problems that are encountered in such a context, little is known about how well they will perform in practice. A goal of this paper is to specifically discuss these approaches in the context of a DDC model with serial correlation and to provide some intuition and evidence about the relative strengths of the various approximation approaches and the situations in which the various approaches are likely to perform well.

The paper proceeds as follows. Section 2 discusses the basic dynamic, discrete choice set-up, the problems that arise when serially correlated continuous variables are present, and a variety of possible approximation approaches to deal with these problems. Section 3 provides some intuition and evidence about the relative performance of the various approximation methods. Section 4 provides an empirical example that accommodates serial correlation, and Section 5 concludes.

2. DYNAMIC, DISCRETE CHOICE WITH SERIALLY CORRELATED VARIABLES

Assume that each individual faces a finite decision horizon ending at time T. At any time t, the individual's objective is to choose an action $a \in A(t)$ that maximizes the expected present value of remaining lifetime rewards.

The current period reward in year t associated with action a is given by

$$u(\epsilon_t, a) + \nu(a) \tag{1}$$

where ν represents a transitory shock to utility that depends on the action that is taken and ϵ_t represents a set of continuous random variables that are potentially serially correlated. In a retirement model, a person's health would be serially correlated and ϵ_t could represent characteristics of a person's health at time t. In a matching model, ϵ_t could be a person's belief at time t about match quality. In the following discussion, we abstract from issues related to the presence of discrete state variables because methods for dealing with these variables are well known.

For notational and illustrative simplicity, we will assume that ϵ_t is independent of ν_t. It is also assumed that the distribution P of ϵ_t depends on only one lagged value

[3] As will be seen, the methods that allow for continuous, serially correlated variables require numerical integration, such as Gaussian quadrature (or simulation), to compute the expected future utility component of value functions. Thus, because these numerical integration methods also involve summations, the general nature of the calculations for the continuous case is similar to that of the discretization case. However, the computational costs will be different for the continuous and discretization methods because the number of elements that must be summed is determined by the number of quadrature points (or simulation draws) in the continuous case and is determined by the number of discretization nodes in the discretization case.

[4] Some of the earliest models that allow serial correlation of some type include those of Pakes (1986), Christenson (1990), and Berkovec and Stern (1991). The last model uses an error structure similar to Miller (1984) to estimate a model in which future realizations of serially correlated unobservables are unknown to the econometrician but are entirely deterministic from the standpoint of the agent in the model. Keane and Wolpin (1997) suggest the benefit of using this type of permanent heterogeneity with a specification of the sort proposed by Heckman and Singer (1984).

$$P(\epsilon_{t+1}|\epsilon_t, \epsilon_{t-1}, \ldots, \epsilon_{t-k+1}) = P(\epsilon_{t+1}|\epsilon_t) \quad \text{for} \quad k = 2, \ldots t \tag{2}$$

Although the discussion in the remainder of the paper is general in nature, in order to fix ideas we will focus primarily on the case where the transition distribution P is normal with a mean that depends on ϵ_t,

$$P(\epsilon_{t+1}|\epsilon_t) = N\big(\Psi(\epsilon_t), \sigma^2\big) \tag{3}$$

We pay close attention to this case in part because the normal distribution is likely to be a popular choice in many applications that involve serial correlation. Examples include empirical implementations of models of Bayesian learning and a wide array of applications in which an autoregressive process would represent a useful relaxation of the standard assumption of serial independence. The illustrative empirical example in Section 4 involves the specification of an AR(1) process for the unobservable in a wage equation.[5]

The econometrician's problem is to compute the expected value of lifetime rewards net of the current period transitory shock, $\nu(a)$, for each alternative, a, that a person considers. As is well known, the expected reward for a particular choice a can be represented recursively by the Bellman equation (Bellman, 1957):

$$v_t(\epsilon, a) = u(\epsilon, a) + \beta \int G(\{v_{t+1}(\epsilon', a'), a' \in A\}) p(\mathrm{d}\epsilon'|\epsilon, a) \tag{4}$$

where β is the discount factor and p is the density associated with equation (3). G is the 'Social Surplus function'

$$G(\{v_{t+1}(\epsilon', a'), a' \in A\}) = \int \cdots \int \max_{a' \in A} [v_{t+1}(\epsilon', a') + \nu(a')] \, g\,(\mathrm{d}\nu(1), \ldots, \mathrm{d}\nu(|A|)) \tag{5}$$

where g is the joint distribution of the transitory shocks to preferences.[6]

In the finite horizon case, the solution of equation (4) typically takes place by backwards recursion starting in the terminal period T. At every time t the goal is to be able to compute $v_t(\epsilon,a)$ for every value of ϵ_t that could enter the choice probabilities at time t or are needed during the recursion in equation (4) to compute the choice-specific value functions in the periods t-1, $t-2, \ldots, 1$. The primary task in this process is the evaluation of the conditional expectation in equation (4). That is, letting $G(\{v_{t+1}(\epsilon',a'),a' \in A\})$ be denoted $W_{t+1}(\epsilon')$, the primary problem of the econometrician is to evaluate conditional expectations of the form

$$EW_{t+1}(\epsilon) = \int_{-\infty}^{\infty} W_{t+1}(\epsilon') p(\mathrm{d}\epsilon'|\epsilon) \tag{6}$$

Note that although it will not always be made explicit in the remainder of the paper, EW_{t+1} is a function of the serially correlated state variable at time t, and W_{t+1} is a function of the serially correlated state variable at time $t+1$.

[5] It is important to note that, while it seems reasonable to believe that the quadrature methods described below will perform well when the random variable ϵ_t has an unbounded support (and this will be examined), in most cases convergence properties have been established under the assumption that the domain is closed and bounded. For example, Tauchen and Hussey (1991) 'follow the tradition in this area of research of applying results deduced for the case of bounded support to models with unbounded support'.

[6] For example, in the popular case where the ν's are IID extreme value, G has a convenient closed-form solution (see e.g. Rust, 1987).

The integral in equation (6) will typically not have a closed-form solution in which case numerical approximation of the integral will be necessary. In order to obtain an approximation $\widehat{EW}_{t+1}(\epsilon)$, the econometrician will need to choose a method of numerical integration. Two prominent possibilities are Gaussian quadrature and Monte Carlo simulation. Assuming for the time being that $W_{t+1}(\epsilon')$ can be computed exactly for all necessary values of ϵ', both numerical integration methods can be written in the form

$$E\hat{W}_{t+1}(\epsilon) = \sum_{i=1}^{S} w_i(\epsilon, \{x_1(\epsilon), \ldots x_S(\epsilon)\}) W_{t+1}(x_i(\epsilon)) \tag{7}$$

For example, the Hermite quadrature formula provides approximations for integrals of the form $\int_{-\infty}^{\infty} Q(x) e^{-x^2} \mathrm{d}x$ that, like other quadrature methods, are exact when Q is a polynomial of degree smaller than $2S$. When P is normal as in equation (3), the integral in equation (6) can be written in this form by applying a simple transformation of variables. Thus, as shown by row 1 of Table I for the case where ϵ is one-dimensional, w_i, $i = 1, \ldots, S$, and x_i, $i = 1, \ldots, S$ are functions of the Hermite quadrature weights w_i^{H}, $i = 1, \ldots, S$, and the Hermite quadrature points x_i^{H}, $i = 1, \ldots, S$ that are published in several books including Stroud and Secrest (1966). As shown in row 7 of Table I, the simulation analogue of the Hermite quadrature approach involves setting $w_i = 1/S$ and determining x_i, $i = 1, \ldots, S$, by random draws from the density $p(\epsilon' | \epsilon)$.

Gauss–Legendre quadrature represents an alternative to Hermite quadrature that provides approximations for integrals of the form $\int_{-1}^{1} Q(x)(1)\mathrm{d}x$, and, as a result, is useful regardless of the distribution P. The integral in equation (6) can be put in this form by first transforming the integral over the real line into an integral over the [0,1] interval using the transformation of variables

$$\int_{-\infty}^{\infty} W_{t+1}(\epsilon') p(\mathrm{d}\epsilon' | \epsilon) = \int_{0}^{1} W_{t+1}\big(P^{-1}(u|\epsilon)\big) \mathrm{d}u \tag{8}$$

and then using a simple transformation of variables which transforms this integral into one over the interval $[-1,1]$.[7] The resulting quadrature formula appears in row 2 of Table I where w_i^{L}, $i = 1, \ldots, S$, and x_i^{L}, $i = 1, \ldots, S$, are the weights and points associated with the Gauss–Legendre method. The simulation analogue appears in row 8.

However, although the previous discussion assumes that W_{t+1} can be computed exactly for all x_i that will arise during the computation of equation (7), in practice this is not the case when the Hermite quadrature, Gauss–Legendre quadrature, or their simulation analogues discussed above are used. The reason stems from the recursive nature of the backwards solution process associated with equations (4)–(6). This process implies that computing $W_{t+1}(x_i)$ for each of the S values of x_i in equation (7) will require knowledge of $v_{t+1}(x_i, a)$ for each possible choice $a \in A$. Computing each of these choice-specific value functions requires knowledge of $EW_{t+2}(x_i)$ which, as seen by rolling equation (7) forward one period, requires evaluating W_{t+2} for S values of ϵ_{t+2} which are determined as functions of x_i. Repetitions of this argument imply that evaluating $EW_{t+1}(\epsilon_t)$ for a single value of ϵ_t requires evaluating W_{t+r} for a number of values of

[7] See Judd (1998) for a description of the second transformation and alternative forms of the first transformation (equation (8)).

Table I. Summary of methods for approximating EW_{t+1}

	x_i, $i = 1,\dots,S$	w_i, $i = 1,\dots,S$
Quadrature		
(1) Interpolating Hermite	$\sqrt{2}\sigma x_i^{\mathrm{H}} + \Psi(\epsilon)$	$\frac{1}{\sqrt{\pi}} w_i^{\mathrm{H}}$
(2) Interpolating Gauss-Legendre	$P^{-1}\left(\frac{x_i^{\mathrm{L}}+1}{2}\vert\epsilon\right)$	$\frac{1}{2} w_i^{\mathrm{L}}$
(3) Self-interpolating Hermite	$\sqrt{2}\sigma x_i^{\mathrm{H}} + \Psi(\mu)$	$\frac{1}{\sqrt{\pi}} w_i^{\mathrm{H}} \frac{p(x_i\vert\epsilon)}{p(x_i\vert\mu)}$
(4) Normalized self-interpolating Hermite	$\sqrt{2}\sigma x_i^{\mathrm{H}} + \Psi(\mu)$	$\frac{w_i^{\mathrm{H}} p(x_i\vert\epsilon)/p(x_i\vert\mu)}{\sum_{j=1}^{S} w_j^{\mathrm{H}} p(x_j\vert\epsilon)/p(x_j\vert\mu)}$
(5) Self-interpolating Gauss–Legendre	$\frac{(x_i^{\mathrm{L}}+1)(b-a)}{2} + a$	$\frac{b-a}{2} w_i^{\mathrm{L}} p(x_i\vert\epsilon)$
(6) Normalized self-interpolating Gauss–Legendre	$\frac{(x_i^{\mathrm{L}}+1)(b-a)}{2} + a$	$w_i^{\mathrm{L}} p(x_i\vert\epsilon)/\sum_{j=1}^{S} w_j^{\mathrm{L}} p(x_j\vert\epsilon)$
Simulation		
(7) Analogue to (1)	Random draws from $N(\Psi(\epsilon),\sigma^2)$	$\frac{1}{S}$
(8) Analogue to (2)	$P^{-1}(u\vert\epsilon)$ for random draws u from uniform [0,1]	$\frac{1}{S}$
(9) Analogue to (3)	Random draws from $N(\mu,\sigma^2)$	$\frac{1}{S}\frac{p(x_i\vert\epsilon)}{p(x_i\vert\mu)}$
(10) Analogue to (4)	Random draws from $N(\mu,\sigma^2)$	$\frac{p(x_i\vert\epsilon)/p(x_i\vert\mu)}{\sum_{j=1}^{S} p(x_j\vert\epsilon)/p(x_j\vert\mu)}$
(11) Analogue to 5	Random draws from uniform [a,b]	$\frac{b-a}{S} p(x_i\vert\epsilon)$
(12) Analogue to 6	Random draws from uniform [a,b]	$\frac{p(x_i\vert\epsilon)}{\sum_{j=1}^{S} p(x_j\vert\epsilon)}$

Note: x_i^{H} and w_i^{H}, $i = 1,\dots,S$ are the points and weights associated with the Hermite quadrature. x_i^{L} and w_i^{L}, $i = 1,\dots,S$ are the points and weights associated with the Gauss–Legendre quadrature. p is the transition density of ϵ_{t+1} given ϵ_t. P is the associated distribution function.

ϵ_{t+r} that is on the order of S^r. While this number meets the backwards recursion solution requirement of being finite, it is typically too large to make computation feasible given current computer resources.

A solution to this problem involves replacing $W_{t+1}(x_i(\epsilon))$ with approximated/interpolated values $\hat{W}_{t+1}(x_i(\epsilon))$ for some or all values of $x_i(\epsilon)$, $i = 1,\dots,S$. There are two general strategies for doing this. One is 'discretization'. In this approach, values of W_{t+1} are 'precomputed' for a set of N predefined ϵ_{t+1} 'grid points', $\{\epsilon^{1*},\dots,\epsilon^{N*}\}$, and these precomputed values are used to approximate/interpolate values of $W_{t+1}(\epsilon')$ for all values of ϵ' that are not on the predefined grid.[8] In this paper, we primarily take the approach of equally spacing the grid points, but this does not have to be the case.[9] The discretization approach for approximating W_{t+1} can be written generally in the form

$$\hat{W}_{t+1}(\epsilon') = \sum_{i=1}^{N} m_i\big(\epsilon', \{\epsilon^{1*}, \dots, \epsilon^{N*}\}\big)\, W_{t+1}(\epsilon^{i*}) \tag{9}$$

where m_i are weights that often sum to one. This class of methods includes various types of interpolation, as well as other non-parametric methods such as local linear regression.[10] This approach with linear interpolation between grid points is taken to allow serial correlation in the empirical example of Section 4 and in Stinebrickner (1996, 2001a) and has also been used more recently in dynamic programming applications by Brien *et al.* (1998, unpublished manuscript) and French (1999, unpublished manuscript).

An alternative strategy for approximating W_{t+1} is a 'parametric approach' in which a parametric family $W_{t+1}(\epsilon',\theta)$ is specified and the $K \times 1$ vector of unknown parameters θ is estimated by regression or some other curve-fitting method using the precomputed values $W_{t+1}(\epsilon^{i*})$, $i = 1,\dots,N$. One common approach is to use linear approximation in which the approximated value $\hat{W}_{t+1}(\epsilon')$ is expressed as a linear combination of K 'basis functions' $\{\Gamma_1(\epsilon'),\dots,\Gamma_K(\epsilon')\}$,

$$\hat{W}_{t+1}(\epsilon') = \sum_{k=1}^{K} \hat{\theta}_k \Gamma_k(\epsilon') \tag{10}$$

This is the approach taken by Keane and Wolpin (1994). Although they discuss several types of specifications for equation (10), the most obvious specification that permits serially correlated ϵ (when ϵ is one-dimensional) would involve the polynomial specification that results from setting $\Gamma_k(\epsilon) = \epsilon^{k-1}$. Note that when ordinary least squares (OLS) is used for estimation, $\hat{\theta}_k$, $k = 1,\dots,K$, will be linear functions of $W_{t+1}(\epsilon^{i*})$, $i = 1,\dots,N$. As a result, in this case a natural connection can be made between the parametric strategy of this type and the discretization strategy because equation (10) can be written using the same general representation as equation (9).

The Hermite quadrature, Gauss–Legendre quadrature, and their simulation analogues described above require the approximation/interpolation of $W_{t+1}(\epsilon')$ because the state space becomes very large as a result of the fact that the x_i's in equation (7) vary with ϵ_t. For the remainder of the paper, we refer to this class as the set of 'interpolating' methods. An alternative 'self-interpolating' approach, which avoids the approximation/interpolation of the W_{t+1} values in equation (7), has been described in economic contexts by Tauchen and Hussey (1991) and Rust (1997). The self-interpolating methods are based on rewriting equation (6)′in the form

$$EW_{t+1}(\epsilon) = \int W_{t+1}(\epsilon')\, \frac{p(\epsilon'|\epsilon)}{g(\epsilon')}\, g(\epsilon')\mathrm{d}\epsilon' \tag{11}$$

[8] The time $t+1$ subscript is suppressed on the grid points. Although not made apparent in the equation below, the precomputed value $W_{t+1}(\epsilon^{i*})$ associated with a grid point ϵ^{i*} will itself not be exact because it typically relies on knowledge of $W_{t+2}(x_j(\epsilon^{i*}))$ for values of x_j that are not grid points at time $t+2$.

[9] For example, the grid points could be determined as a random sample from an appropriate distribution.

[10] See Härdle and Linton (1994) for a description of non parametric methods.

and determining the points x_i in equation (7) on the basis of an 'importance' density $g(\epsilon')$. For example, when P is normal, Tauchen and Hussey propose letting $g(\epsilon') = p(\epsilon' \mid \mu)$ where μ is the unconditional mean of the ϵ process. Using this weighting function and assuming that ϵ_{t+1} is normal as in equation (3) leads to the self-interpolating Hermite quadrature method in row 3 of Table I and its simulation analogue in row 9. A variation of these methods involves normalizing the weights in row 3 and row 9 so that they sum to one. These normalized versions are shown in row 4 and row 10 of Table I.

The self-interpolating nature of the methods in rows 3, 4, 9, and 10 stems from the fact that the density $g(\epsilon_{t+1})$ that is used to determine the points x_i , $i = 1, \ldots, S$, does not depend on ϵ_t. This implies that the points x_i in equation (7) do not vary with ϵ_t. Thus, W_{t+1} must only be solved at time t for the S values of ϵ_{t+1} given by x_i, $i = 1, \ldots, S$, in Table I, and, as a result, interpolation of W_{t+1} is not necessary. As can be seen from rows 3, 4, 9, and 10 of Table I, ϵ_t now influences $E\hat{W}_{t+1}$ through the weights w_i, $i = 1, \ldots, S$, which are functions of $p(x_i \mid \epsilon_t)$. However, the presence of weights that vary with ϵ_t does not necessitate interpolation because $P(\epsilon_{t+1} \mid \epsilon_t)$ is well defined and can be computed easily for all values of ϵ_t.

Another self-interpolating option, that is applicable regardless of the distribution P, is simply to let $g(\epsilon') = 1$ in which case the appropriate quadrature formula is the Gauss–Legendre formula. The transformation in equation (8) is not useful from the standpoint of designing a self-interpolating method because it implies that the values of ϵ_{t+1} at which W_{t+1} is evaluated depend on ϵ_t. However, because

$$\int_{-\infty}^{\infty} W_{t+1}(\epsilon') \, \frac{p(\epsilon'|\epsilon)}{g(\epsilon')} \, g(\epsilon')\mathrm{d}\epsilon = \lim_{a \to -\infty, b \to \infty} \int_a^b W_{t+1}(\epsilon') \, \frac{p(\epsilon'|\epsilon)}{g(\epsilon')} \, g(\epsilon')\mathrm{d}\epsilon$$

whenever the limit exists, a direct approach to the approximation can be taken by simply choosing a and b to be finite but 'relatively' large. Methods of this type are shown in rows 5, 6, 11, and 12 of Table I. Unfortunately, as suggested by Judd(1998), this truncation approach is likely to be slow.

3. APPROXIMATION QUALITY COMPARISONS

In this section the relative performance of the various methods from Section 2 is examined. In order to keep the discussion tractable, the focus is on the case where ϵ is one-dimensional. It is important to note in advance that the specification of W_{t+1} and P is likely to play an important role in determining how the various approximation methods perform. W_{t+1} and P can potentially vary significantly across applications and given enough exploration, one could almost certainly find specifications such that any particular method performed very well (or very poorly) relative to the others. Because it is not possible to perform testing that is exhaustive over the entire space of W_{t+1}, P specifications, it is not the intention of this paper to conclude that a particular method(s) is superior to others. Rather, the hope is to highlight how certain characteristics of W_{t+1} or P may influence the relative performance of the methods. Hopefully, this will increase the chances that a researcher will be able to choose a method that is useful in his or her application and avoid methods that may perform poorly.

Designing an experiment that allows comparisons of $E\hat{W}_{t+1}(\epsilon_t)$ to $EW_{t+1}(\epsilon_t)$ raises some difficulties when P is normal because finding functional forms of $W_{t+1}(\epsilon_{t+1})$ for which $EW_{t+1}(\epsilon_t)$ has a closed-form solution is not easy. The approach taken below is as follows. First, approximation quality is examined under a baseline specification in which W_{t+1} is specified as the polynomial of degree thirteen in ϵ_{t+1} that is shown in Figure 1, and P is described by the AR(1) process

$$\epsilon_{t+1} = \rho\, \epsilon_t + e_{t+1}, e_{t+1} \sim \mathrm{N}(0, \sigma_e^2) \tag{12}$$

with $\sigma_e = 0.30$ and $\rho = 0.95$.[11] For this specification, we then examine how relative approximation performance changes as the transition density p changes as a result of changing ρ and as the polynomial specification of W_{t+1} changes in several simple ways. For the baseline specification and the modifications of the baseline specification, W is a polynomial and the desired comparisons are possible because the true value of $EW_{t+1}(\epsilon_t)$ can be uncovered for any ϵ_t using the Hermite quadrature approximation described in the paragraph after equation (7) given enough quadrature points ($S \geq 7$) and given exact evaluation of $W_{t+1}(x_i(\epsilon))$, $i = 1, \ldots, S$. It is important to stress that this choice of W_{t+1} will be informative despite the fact that $EW_{t+1}(\epsilon_t)$ can be computed exactly as described above, because, as discussed in Section 2, none of the approximation methods that allow the estimation of DDC models with serial correlation are equivalent to the Hermite quadrature method with exact evaluation of W_{t+1}.[12]

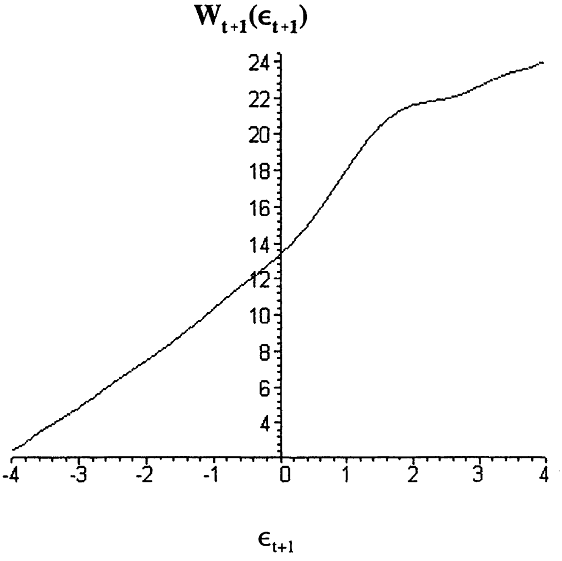
Before moving to the approximation comparisons, it is necessary to establish a concept of computational cost which can be held constant while comparing approximation quality across methods. For each value of ϵ_t for which value functions are solved by backwards recursion, computing $\hat{EW}_{t+1}(\epsilon_t)$ involves summing the S terms in equation (7). Therefore, the total number of terms that must be summed at time t of the backwards recursion process is the product of S and the number of values of ϵ_t for which value functions are being solved by backwards recursion at t. We refer to this total number as C and use it as a measure of total computational cost.[13] For the self-interpolating methods in Table I, the number of values of ϵ_t for which value functions need to be solved during the backwards recursion process is also S, which implies that $C = S^*S$. However, for the interpolating methods, the number of grid points, N, is chosen separately from S. Thus, $C = N^*S$ and for any particular value of C there will exist multiple combinations of N and S that yield this cost. As will be seen, the optimal choice of S for the interpolating methods will typically be different from $\sqrt{C}$. When showing the approximation

[11] Although it is necessary to establish a starting point, the baseline specification of W is somewhat arbitrary and this motivates the necessity of examining approximation when W is changed. In this case, the chosen form was motivated in part by examining the shape of W in the empirical example of Section 4.

[12] The interpolating Hermite quadrature method differs both because it requires the use of approximated values of W_{t+1} and because it will turn out that we will often be interested in approximation quality for values $S < 7$. In this case there may exist polynomials of degree $2S-1$ that are 'close' to W_{t+1} but they will not be identical. Note that this situation is essentially identical to the case in which W_{t+1} is not a polynomial because in this case, as S increases, the Weistrauss theorem ensures that there will exist polynomials of degree $2S-1$ which become 'close' although not identical to the non-polynomial W_{t+1} (at least on a closed and bounded interval). As a result, the use of polynomial specification for W_{t+1} is not likely to be particularly limiting. The interpolating methods with Gauss–Legendre quadrature differ for the reasons above and because $W_{t+1}(P^{-1})$ is not typically a polynomial even if W_{t+1} is a polynomial. Similarly, the self-interpolating methods involve the evaluation of $W_{t+1}p/g$ and $W_{t+1}p$ rather than the polynomial W_{t+1}.

Approximation quality was also examined for several entirely different, non-polynomial specifications of W_{t+1}. For these cases, comparisons were possible because it was possible to compute an error bound for the value that we use for the 'truth'. The conclusions for these specifications were consistent with those found with the polynomial specifications, but are not included due to space limitations.

[13] The computational burden of increasing C would clearly depend on the rest of the model. For example, if many other state variables are present, doubling estimation time could be very burdensome. In other cases, it may not be problematic.



Note: $W_{t+1}(\epsilon_{t+1})$=13.71+3.60 ϵ_{t+1}+1.30 ϵ_{t+1}^{2}+.64 ϵ_{t+1}^{3}-.47 $\epsilon_{t+1}^{4\cdot}$-.35ϵ_{t+1}^{5}+.07 ϵ_{t+1}^{6} +.06ϵ_{t+1}^{7}-.005 ϵ_{t+1}^{8}-.005 ϵ_{t+1}^{9}+.0001 ϵ_{t+1}^{10}+.0002 ϵ_{t+1}^{11}-.000002 ϵ_{t+1}^{12}-.0000004ϵ_{t+1}^{13}

Figure 1(a)

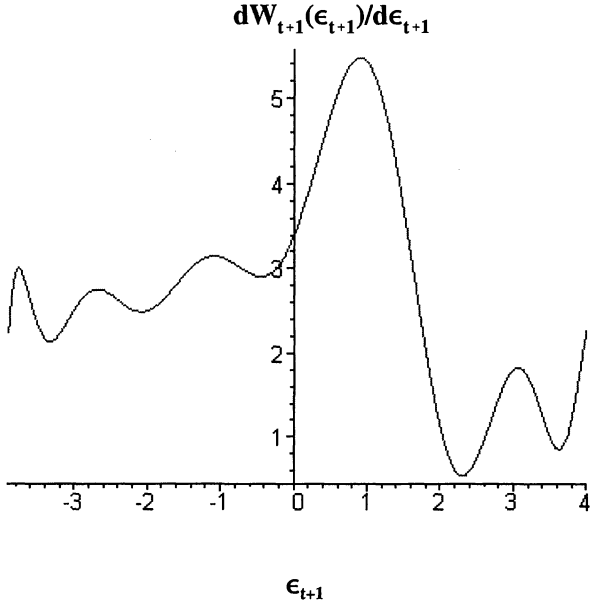


Figure 1(b)

quality for the interpolating methods, the following analysis uses the combination of N and S that yields the best approximation quality. Thus, the numbers associated with the interpolating methods should be thought of as the best feasible approximation quality given a particular computational cost. Note that is important to keep in mind that C does not include the computational costs of approximating/interpolating W_{t+1} for the interpolating methods. These costs are small in the single-dimension case we are examining, but would become more substantial in higher dimensions.

In order to provide the approximated values $\hat{W}_{t+1}(\epsilon_{t+1})$ that are required for the interpolating methods to compute $E\hat{W}_{t+1}(\epsilon_t)$, the 'parametric' approach requires a specification of the parametric family $W_{t+1}(\epsilon_{t+1}, \theta)$ and the 'discretization' approach requires the choice of a non-parametric method for use with the N grid points. To make the following exposition tractable, we begin by computing the interpolating methods using a discretization approach with $\hat{W}_{t+1}(\epsilon_{t+1})$ determined by a linear interpolation between the N grid points. This is the approach used in the empirical example of Section 4. A prominent alternative is the parametric approach used by Keane and Wolpin (1994) in which θ is estimated by OLS. As discussed in Section 2, a natural connection between these methods can be made. A closer examination of this relationship in the context of the tests described below is examined in the Appendix.

3.1 Relative Approximation Performance and ϵ_t

Recall from Section 2 and Table I that the interpolating Hermite quadrature method (and its simulation analogue) involves the evaluation of $W_{t+1}(\epsilon_{t+1})$, the interpolating Gauss–Legendre quadrature method (and its simulation analog) involves the evaluation of $W_{t+1}(P^{-1}(u \mid \epsilon_t))$, the self-interpolating Hermite methods (and their simulation analogues) involve the evaluation of $W_{t+1}(\epsilon_{t+1})p(\epsilon_{t+1} \mid \epsilon_t)/g(\epsilon_{t+1})$, and the self-interpolating Gauss–Legendre methods (and their simulation analogues) involve the evaluation of $W_{t+1}(\epsilon_{t+1})p(\epsilon_{t+1} \mid \epsilon_t)$. The performance of both the quadrature methods and the simulation methods depends on the shape of these functions. As discussed earlier, S point quadrature formulas provide exact solutions if the relevant function (called Q in Section 2) is a polynomial of degree less than $2S$. Intuitively, this implies that quadrature methods will work well with relatively small S if the relevant function can be approximated well by a polynomial of low degree. Similarly, the simulation methods will tend to perform well when the variance of the function that is being evaluated is small. Thus, differences in the shapes of the functions being evaluated can generate differences in approximation quality across methods.

Further, because the shape of the function being evaluated for a particular method may vary with ϵ_t, the relative approximation quality of the various methods may vary significantly with ϵ_t. $W_{t+1}(\epsilon_{t+1})$ does not depend on ϵ_t and appears as in Figure 1(a) for any ϵ_t. Figures 2(a) and 2(b), which show $W_{t+1}(P^{-1}(u \mid \epsilon_t = 0.02))$ and $W_{t+1}(P^{-1}(u \mid \epsilon_t = 2.25))$ for $\rho = 0.95$, indicate that the shape of $W_{t+1}(P^{-1}(u \mid \epsilon_t))$ does not vary substantially with ϵ_t and is reasonably similar in nature to $W_{t+1}(\epsilon_{t+1})$. Thus, it seems reasonable to believe that the two interpolating methods may perform similarly. Figures 3(a) and 3(b) show $W_{t+1}(\epsilon_{t+1})p(\epsilon_{t+1} \mid \epsilon_t = 0.02)/p(\epsilon_{t+1} \mid \mu = 0)$ and $W_{t+1}(\epsilon_{t+1})p(\epsilon_{t+1} \mid \epsilon_t = 2.25)/p(\epsilon_{t+1} \mid \mu = 0)$. For $\epsilon_t = 0.02$ (and other values of ϵ_t that are close to μ), the self-interpolating Hermite quadrature methods (and their simulation analogues) should be expected to perform better than the interpolating methods (and their simulation analogues) because, as shown by Figure 3(a), $W_{t+1}p/g$ is very close in shape to W_{t+1}, but,

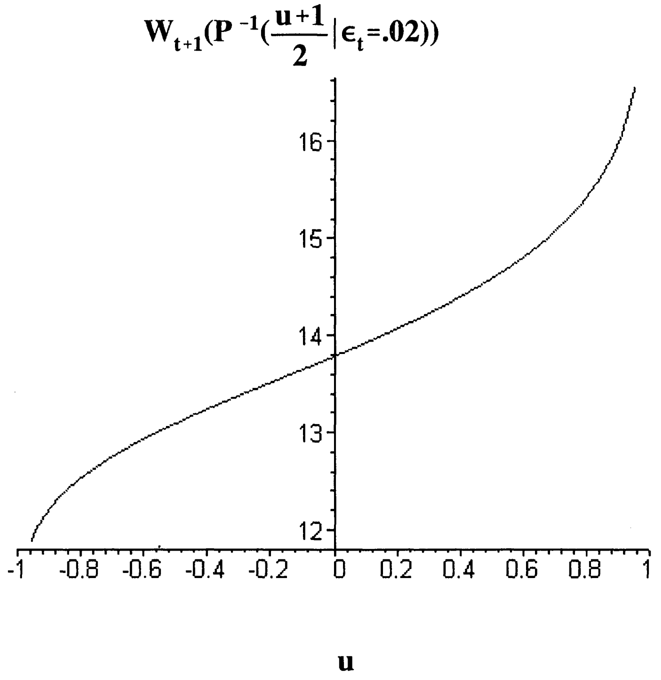


Figure 2(a)

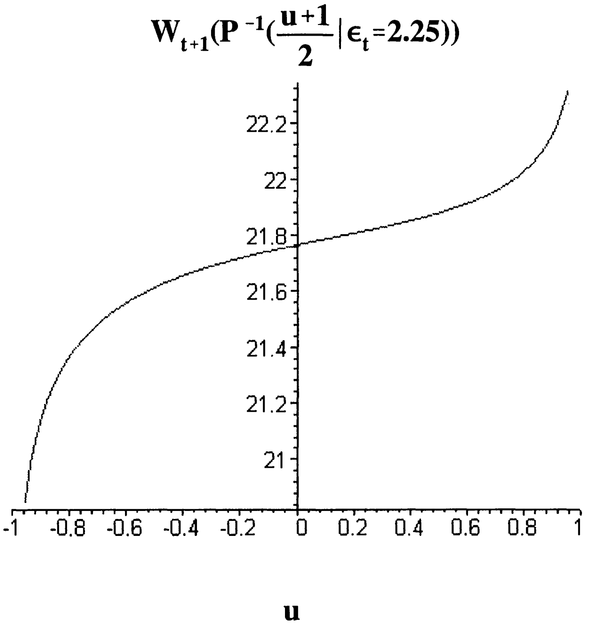


Figure 2(b)

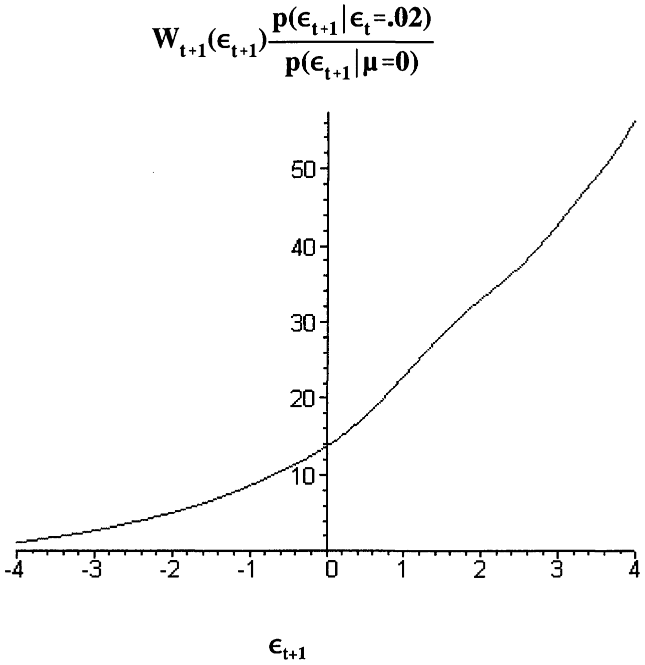


Figure 3(a)

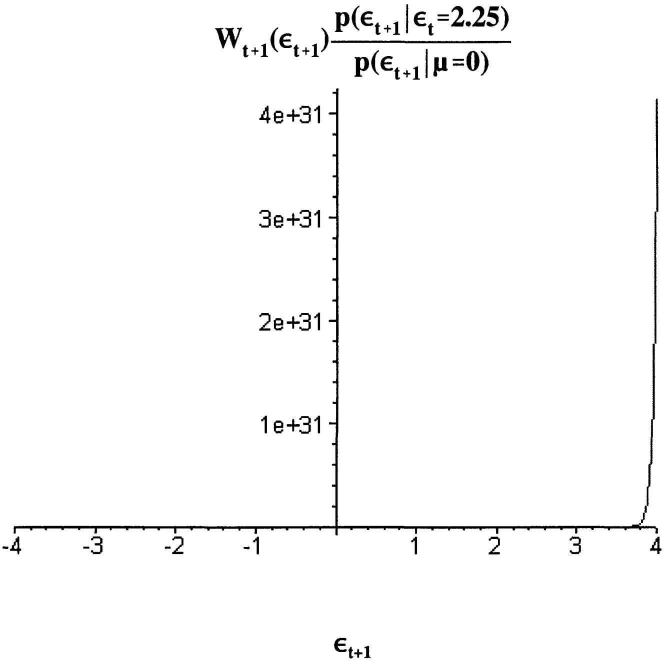


Figure 3(b)

unlike the interpolating methods, the self-interpolating Hermite methods do not require the interpolation of the $W_{t+1}(\epsilon_{t+1})$ values. However, for $\epsilon_t = 2.25$ (and other values of ϵ_t that are further from μ), Figure 3(b) indicates that a low-order polynomial would not approximate $W_{t+1}p/g$ well and we should expect the self-interpolating Hermite methods to perform worse than the interpolating methods as long as the error associated with approximating $W_{t+1}(\epsilon_{t+1})$ for the interpolating methods is not extremely large. Figures 4(a) and 4(b) show that, although $W_{t+1}(\epsilon_{t+1})p(\epsilon_{t+1}|\epsilon_t)$ is not likely to be approximated well by a polynomial of low order, the shape of the function $W_{t+1}p$ does not change dramatically with ϵ_t. Thus, we might expect that the self-interpolating Gauss–Legendre methods will perform similarly for all ϵ_t but will perhaps perform generally worse than the other methods.

This intuition is confirmed by Table II which shows how $|E\hat{W}_{t+1}(\epsilon_t) - EW_{t+1}(\epsilon_t)|$ varies with ϵ_t for the quadrature methods in Table I for $C = 49$, where W is defined as in Figure 1, $\rho = 0.95$, and, as in all of the other tests in this paper, $\sigma_e = 0.30$. We return to the approximation quality of the simulation methods shortly, but do not include them in Table II because they only reinforce the conclusions obtained by looking at the quadrature methods. In order to determine the types of values of ϵ_t that are 'likely' to appear, we take into account that the unconditional distribution of ϵ_t at time t has a limiting distribution (as t becomes large) of

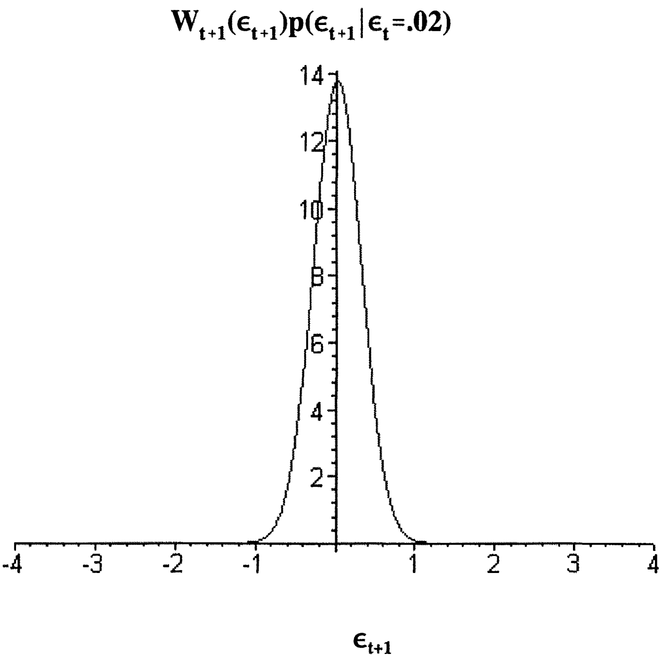
$$H(\epsilon_t) = N\left(\frac{\mu}{1-\rho}, \frac{\sigma_e^2}{1-\rho^2}\right) \tag{13}$$

where μ is the unconditional mean of the ϵ process.

Without loss of generality, for the remainder of the paper we set $\mu = 0$. Table II shows that the Hermite and Gauss–Legendre interpolating methods (methods 1 and 2) yield similar approximations and approximation quality is roughly constant across values of ϵ_t for each method. The self-interpolating Hermite methods (methods 3 and 4) perform much worse than the interpolating methods for ϵ_t which are large in absolute value, but perform very well for values of ϵ_t that are close to $\mu = 0$. The self-interpolating Gauss–Legendre methods (methods 5 and 6) perform quite poorly although better than the self-interpolating Hermite methods for values of ϵ_t that are far from μ.

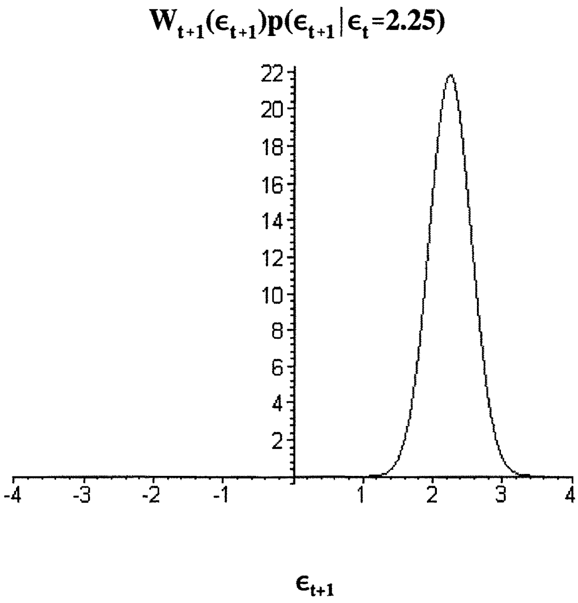
To understand more fully the problems that arise with the self-interpolating Hermite methods when ϵ_t is not close to μ, consider the weights w_i in row 3 of Table I that are used to compute the non-normalized self-interpolating Hermite quadrature. When ϵ_t is not close to μ, $p(\epsilon_{t+1}|\epsilon_t)/p(\epsilon_{t+1}|\mu)$ becomes very small and the weights w_i are small. For example, when $S = 7$ ($C = 49$) and $\epsilon_t = 2.25$, the largest of the seven weights is only 0.0002 and the next largest is 6.13×10^{-6}. This leads to the approximation $E\hat{W}_{t+1}(2.25) = 0.0402$ which substantially misses the truth of $EW_{t+1}(2.25) = 21.7217512$. This problem is corrected asymptotically because, for large enough S, x_i^{H} in row 3 of Table I becomes large for some i and W_{t+1} is evaluated at values of x_i for which $p(x_i|\epsilon)/p(x_i|\mu)$ is large; at $S = 40$, $E\hat{W}_{t+1}(2.25) = 21.721732$. Nonetheless, until S becomes quite large the method performs quite poorly. At $S = 10$, $E\hat{W}_{t+1}(2.25) = 0.9328$ and at $S = 20$, $E\hat{W}_{t+1}(2.25) = 19.5969$.

The self-interpolating normalized Hermite performs substantially better than the non-normalized version for ϵ_t far from μ. For example, for $S = 7$, the normalized version produces $E\hat{W}_{t+1}(2.25) = 19.211$. The improvement stems from the fact that, because the weights in row 4 of Table I sum to one, the approximated value $E\hat{W}_{t+1}$ is a weighted average of actual W_{t+1} values. Nonetheless, the normalized self-interpolating case still understates the truth



$W_{t+1}(\epsilon_{t+1})p(\epsilon_{t+1}|\epsilon_t=.02)$

Figure 4(a)



$W_{t+1}(\epsilon_{t+1})p(\epsilon_{t+1}|\epsilon_t=2.25)$

Figure 4(b)

Table II. Approximation quality for different values of ϵ_t

ϵ_t Method — Table I	−2.25	−1.5	−0.75	−0.25	−0.02	0.00	0.02	0.25	0.75	1.5	2.25
(1)	1.28-3	7.80-3	3.97-3	4.65-3	5.36-2	5.39-2	5.34-2	1.88-2	2.70-2	9.18-2	2.31-3
(2)	1.44-3	4.78-3	1.81-3	1.06-2	2.09-2	2.13-2	2.19-2	1.93-2	1.02-2	4.21-2	9.89-3
(3)	7.35	5.13	1.28-1	1.11-7	1.33-9	0	3.26-9	1.45-6	3.55-1	1.25+1	2.16+1
(4)	2.88	1.00	5.20-2	5.94-7	2.08-9	1.97-10	2.51- 9	8.94-7	6.78-2	1.23	2.51
(5)	4.03	7.30	7.95	4.27	1.30+1	1.30+1	1.29+12	.14	1.14+1	1.64+1	1.11+1
(6)	5.05-1	4.01-2	1.30	8.78-1	3.61-2	1.06-1	1.78 -1	1.25	2.03	2.36-1	3.88-2

Note: The table shows $|E\hat{W}_{t+1}(\epsilon_t) - EW_{t+1}(\epsilon_t)|$ for various values of ϵ_t and is constructed with W as in Figure 1, $\sigma_e = 0.3$, and $\rho = 0.95$ using $C = 49$. All numbers are shown in scientific notation. For example, 1.28-3 = 1.28 × 10-3.

substantially for $S = 7$. This occurs because this method involves a weighted average of W_{t+1} evaluated at seven values of x_i, −1.125, −0.710, −0.346, 0.0, 0.346, 0.710, and 1.125, that are all substantially smaller than the types of values that might be expected given the transition distribution $P(\epsilon_{t+1}|\epsilon_t = 2\ .25)$.[14] This problem is corrected asymptotically because, for large S, x_i^{H} becomes large for some i, W_{t+1} is evaluated at some values of x_i that are much closer to (or larger than) $E(\epsilon_{t+1}|\epsilon_t = 2.25)$ and the weights associated with these values of x_i are high. As evidence of this, at $S = 40$, the normalized self-interpolating yields a value of 21.7217511. However, the method can potentially perform somewhat poorly for values of ϵ_t far from μ until S becomes quite large. At $S = 10$ and $S = 20$, the approximation yields 20.5822 and 21.6610 respectively.

3.2 Summary Measures of Approximation Quality

Because the relative performance of the various methods varies substantially with ϵ_t, it is desirable for comparison purposes to compute a single statistic that takes into account approximation quality across all possible values of ϵ_t. For each of the approximation methods, we consider the average amount that $E\hat{W}_{t+1}(\epsilon_t)$ differs from the truth $EW_{t+1}(\epsilon_t)$:

$$EEW_{t+1} = E\big(|E\hat{W}_{t+1}(\epsilon_t) - EW_{t+1}(\epsilon_t)|\big) = \int_{-\infty}^{\infty} \big(|E\hat{W}_{t+1}(\epsilon_t) - EW_{t+1}(\epsilon_t)|\big)K(\epsilon_t)\mathrm{d}\epsilon_t \qquad (14)$$

where the distribution $K(\epsilon_t)$ serves as a weighting function which determines the importance that is assigned to the approximation quality associated with the various possible values of ϵ_t. Equation (14) is simulated by

$$E\hat{EW}_{t+1} = \frac{1}{6000}\sum_{i=1}^{N} |E\hat{W}_{t+1}(\epsilon_t^i) - EW_{t+1}(\epsilon_t^i)| \qquad (15)$$

[14] This implies that $E\hat{W}_{t+1}$ will necessarily understate the truth in this case because W is increasing in ϵ_t. In this case, $W_{t+1}(1.125) = 19.211$ is assigned a weight of 0.998 in the weighted average. In essence, the other points contribute very little to the approximation.

where ϵ_t^i is the ith of 6000 draws from the distribution $K(\epsilon_t)$. The number of draws is large enough to ensure that the difference between the $E\hat{EW}_{t+1}$ values associated with any two approximation methods is at least ten times larger (and typically much more than ten times larger) than the larger of the standard errors associated with the two simulation estimators. As a result, most of the comparisons in the remainder of the paper are made without reference to sampling variation from the simulation of equation (15).

In order to choose an informative weighting distribution K, it is important to keep in mind that the ultimate purpose of the $E\hat{W}_{t+1}(\epsilon_t)$ approximation is to allow the computation of the choice specific value functions in equation (4) that serve as inputs into the choice probabilities that are needed for estimation. A reasonable way to begin is to set K equal to the distribution H from equation (13). This approach takes into account that values of ϵ_t that are close to μ are more likely to appear in the likelihood function than other values of ϵ_t.[15] It may also be informative to let K be a uniform distribution so that equal weight is put on the approximation quality associated with the various possible values of ϵ_t.[16] For example, suppose that ϵ_t represents health status in a retirement model and that retirement decisions are almost exclusively driven by the onset of very poor health. Then a researcher may be as interested in the approximation quality for values of ϵ_t that represent very poor health as he or she is in the approximation quality for values of ϵ_t that represent average health even if the former are much less likely to occur given the unconditional distribution of health. A comparison of $E\hat{EW}_{t+1}$ computed with a uniform K to $E\hat{EW}_{t+1}$ computed with a normal K provides insight into whether a particular method is performing differently for values of ϵ_t that are far from μ than it is for values of ϵ_t that are close to μ.

The first six rows of Table III show $E\hat{EW}_{t+1}$ values for the quadrature methods in Table I under the baseline specification ($\rho = 95$, $\sigma_e = .30$, and W as in Figure 1) for $C = 9$, 25, 49, 81, and 400.[17] In all of Table III, the two numbers in the first row of a particular box are the approximation values when the normal and uniform distributions are used respectively for the weighting function K. Among the quadrature methods, the interpolating Hermite and Gauss–Legendre methods have similar $E\hat{EW}_{t+1}$ values. The interpolating methods perform substantially better than the self-interpolating methods for all C that were tested and the performance of the self-interpolating Gauss–Legendre method is generally worse than all of the other methods. Table III also shows that the value of S at which the interpolating methods perform best is substantially lower than the value of S that is being used for the self-interpolating methods with the same C. For example, at $C = 49$, the self-interpolating methods use $S = 7$. The interpolating methods could also use $S = 7$ if the number of grid points, N, is also

[15] In particular, for computational reasons ϵ_t^i is drawn from a truncated version of H where the truncation points $\epsilon_t^{\mathrm{Min}}$ are determined by $H(\epsilon_t^{\mathrm{Min}}) = 0.005$ and $H(\epsilon_t^{\mathrm{Max}}) = 0.995$. The results below are robust to the truncation level that was chosen.

[16] $[\epsilon_t^{\mathrm{Min}}, \epsilon_t^{\mathrm{Max}}]$ is used as the support of the uniform distribution.

[17] For the interpolating methods, grid points are spaced evenly over the possible range of values. For example, from row 1 of Table I the range for the interpolating Hermite method can be seen to be $[\sqrt{2}\sigma_e x_1^{\mathrm{H}} + \rho\epsilon_t^{\mathrm{Min}}, \sqrt{2}\sigma_e x_S^{\mathrm{H}} + \rho\epsilon_t^{\mathrm{Max}}]$. For the self-interpolating Gauss–Legendre methods, the truncation levels a and b were chosen by comparing approximation performance at several different levels. It was found that choosing a and b so that $P(a \mid \epsilon_t^{\mathrm{Min}}) = 0.005$ and $P(b \mid \epsilon_t^{\mathrm{Max}}) = 0.995$ generally worked well relative to other choices.

As discussed earlier, sampling variation associated with $E\hat{EW}_{t+1}$ values that arises due to the simulation with 6000 simulation draws is generally small relative to $E\hat{EW}_{t+1}$ and is not included in Table III. For $C = 81$ and normal K, the standard errors associated with $E\hat{EW}_{t+1}$ due to this source for the twelve methods are 2.81e-5, 1.30e-5, 4.65e-2, 4.94e-3, 2.25e-2, 3.90e-3, 7.35e-4, 6.82e-4, 6.51e-1, 1.99e-2, 7.55e-2, 1.20e-2.

Table III. Approximation quality baseline specification ($\rho = 0.95$)

	$C = 9$		$C = 25$		$C = 49$		$C = 81$		$C = 400$		$C = 10{,}000$	
K	$K{\sim}N$	$K{\sim}U$	$K{\sim}N$	$K{\sim}U$	$K{\sim}N$	$K{\sim}U$	$K{\sim}N$	$K{\sim}U$	$K{\sim}N$	$K{\sim}U$	$K{\sim}N$	$K{\sim}U$
Method — Table I												
Quadrature												
(1)	5.97-1	7.97-1	7.19-2	7.54-2	2.05-2	2.08-2	5.83-3	5.78-3	3.95-4	4.15-4		
	$S = 3$		$S = 3$		$S = 3$		$S = 3$		$S = 3$			
(2)	5.69-1	7.50-1	5.85-2	6.04-1	1.00-2	1.01-2	5.05-3	4.99-3	3.19-3	3.23-3		
	$S = 3$		$S = 3$		$S = 3$		$S = 3$		$S = 5$			
(3)	5.68	9.66	3.41	7.82	2.07	6.29	1.26	4.92	3.61-2	3.71-1		
(4)	1.25	2.63	6.28-1	1.66	3.10-1	1.01	1.51-1	6.07-1	2.86-3	2.98-2		
(5)	1.60+1	1.36+1	1.00+1	9.23	5.98	5.19	3.12	2.61	8.62-3	8.20-3		
(6)	1.88	1.70	1.26	1.01	8.29-1	6.50-1	4.93-1	3.78-1	4.65-3	2.33-3		
Simulation												
(7)	7.54-1	9.88-1	1.97-1	1.82-1	1.06-1	1.11-1	4.90-2	4.97-2	2.35-2	2.42-2	9.55-3	(9.96-3)
	(1.01-1)	(7.88-2)	1(1.24-1)	(1.12-1)	(4.33-2)	(4.62-2)	(5.09-2)	(5.24-2)	(2.07-2)	(2.16-2)	(8.19-3)	(8.41-3)
	$S = 3$		$S = 5$		$S = 6$		$S = 6$		$S = 20$		$S = 100$	
(8)	7.45-1	9.81-2	1.98-1	1.81-1	1.18-2	1.17-2	7.19-2	6.58-2	4.28-2	4.28-2	3.80-2	3.36-2
	(9.06-2)	(6.84-2)	(1.21-1)	(1.06-1)	(4.74-2)	(4.77-2)	(4.84-2)	(4.80-2)	(1.80-2)	(1.64-2)	(4.78-3)	(3.65-3)
	$S = 3$		$S = 5$		$S = 6$		$S = 6$		$S = 20$		$S = 100$	
(9)	1.07+1	1.28+1	1.03+1	1.26+1	9.47	1.20+1	1.17+1	1.42+1	1.24+1	1.53+1	6.60	1.02+1
	(1.12+1)	(7.92)	(9.59)	(6.75)	(8.49)	(5.93)	(2.73+1)	(2.52+1)	(3.46+1)	(3.27+1)	(6.07)	(5.37)
(10)	1.95	3.39	1.77	3.16	1.66	3.03	1.59	2.93	1.42	2.69	1.02	2.14
	(4.11-1)	(5.29-1)	(4.18-1)	(5.44-1)	(4.10-1)	(5.35-1)	(3.98-1)	(5.31-1)	(3.49-1)	(4.91-1)	(2.75-1)	(4.06-1)
(11)	1.56+1	1.52+1	1.21+1	1.17+1	1.02+1	9.63	8.73	8.27	5.48	5.44	2.50	2.41
	(4.19)	(2.76)	4.05)	(3.16)	(3.81)	(3.01)	(3.31)	(2.65)	(2.64)	(2.25)	(1.12)	(9.39-1)
(12)	2.86	2.42	1.97	1.60	1.45	1.15	1.14	8.98-1	4.91-1	4.09-1	1.56-1	1.37-1
	(1.44)	(8.60-1)	(1.29)	(7.94-1)	(1.02)	(6.20-1)	(8.99-1)	(5.39-1)	(4.83-1)	(2.96-1)	(6.26-2)	(4.74-2)

Note: The table shows approximation quality for specification with W as in Figure 1, $\sigma_e = 0.3$, and $\rho = 0.95$. The twelve methods are defined as in Table I. For quadrature methods, the first row in each box shows $EE\hat{W}_{t+1}$ in equation (15) for Normal K and Uniform K. For simulation methods, the first row in each box shows average $EE\hat{W}_{t+1}$ value over 100 different sets of simulation draws for x_i's in equation (7). The second row shows standard deviation of $EE\hat{W}_{t+1}$ value over 100 different sets of simulation draws for x_i's in equation (7). Rows associated with interpolating methods show S at which best approximation quality takes place. For self-interpolating methods, S is by definition $C^{0.5}$. All numbers are shown in scientific notation. For example, 5.97-1 = 5.97 × 10-1. True EEW_{t+1} is 14.24 for normal K and 14.45 for uniform K.

chosen to be seven. However, approximation quality is improved by using smaller S. This allows an increase in N which improves the quality of the $\hat{W}_{t+1}(\epsilon_{t+1})$ values. The best approximation performance was found at $S = 3$.[18]

For the quadrature methods, $|E\hat{W}_{t+1}(\epsilon_t^i) - EW_{t+1}(\epsilon_t^i)|$ in equation (15) is deterministic given ϵ_t^i. This is not the case for the simulations methods because the x_i's in equation (7) are random draws from the distributions described in rows 7-12 of Table I. Thus, for the simulation methods we compute the average value of $EE\hat{W}_{t+1}$ over 100 sets of simulation draws for the x_i's. For the simulation methods, the two rows in each box of Table III show the average value of $EE\hat{W}_{t+1}$

[18] For the self-interpolating Hermite quadrature method at $C = 49$, $E\hat{E}W_{t+1}$ with normal K is 0.015, 0.010, 0.035, 0.042, 0.072, and 0.11 at $S = 2$, 3, 4, 5, 6, and 7 respectively.

over the 100 sets of simulation draws and the standard deviation of $E\hat{E}W_{t+1}$ over the 100 sets of simulation draws. The methods based on simulation generally perform poorly relative to the analogous quadrature-based methods. The non-normalized self-interpolating methods appear to be particularly problematic. The reason is that some of the 100 simulation repetitions lead to large outliers for the $E\hat{E}W_{t+1}$ value. Evidence of this is suggested by the large variance in $E\hat{E}W_{t+1}$ that occurs over the 100 simulation repetitions in rows 9 and 11. Although the normalized self-interpolating versions largely avoid this problem, Table III suggests that one should also be careful about using these methods with large ρ. However, it is important to stress that the benefit of simulation methods is likely to reveal itself in problems of higher dimensions. As a result, it is not the primary goal to examine the relationship between the simulation and quadrature based methods, and in the remainder of the paper we concentrate primarily on the quadrature methods. Given the poor performance of the non-normalized methods, we also choose to concentrate primarily on the normalized methods.

3.3 Changes to Transition Distribution *P* — changes in ρ

The performance of the self-interpolating Hermite method should be expected to improve relative to the interpolating methods as ρ is decreased. There are two reasons for this. First, equation (13) indicates that, on average, ϵ_t is closer to μ when ρ is small. For example, when $\rho = 0.95$, $\Pr(-2.507 < \epsilon_t < 2.507) = 0.995$, and when $\rho = 0.75$, $\Pr(-1.183 < \epsilon_t < 1.183) = 0.995$. Second, for a particular ϵ_t, the performance of the self-interpolating Hermite will tend to be better relative to the interpolating methods when ρ is small. This occurs because $W_{t+1}p/g$ can be more easily approximated by a lower degree polynomial when ρ is small. To see this, realize that as $\rho \rightarrow 0$, $W_{t+1}p/g \rightarrow W_{t+1}$ and the self-interpolating method will perform better for any ϵ_t because, unlike the interpolating method, it does not require approximation of the W_{t+1} values that are needed in equation (7).

This intuition is confirmed by Table IV which, for $\rho = 0.75$ and W_{t+1} as in Figure 1(a), shows $E\hat{E}W_{t+1}$ values for the quadrature methods in rows 1, 2, 4, and 6 of Table I. All of the methods perform better than in Table III. The self-interpolating Hermite (method 4) now performs generally better than the interpolating methods (methods 1 and 2) when the normal weighting function is used. However, the interpolating methods remain better except for the

Table IV. Approximation quality baseline specification with $\rho = 0.75$

	$C = 9$		$C = 25$		$C = 49$		$C = 81$		$C = 400$	
K	$K\sim N$	$K\sim U$	$K\sim N$	$K\sim U$	$K\sim N$	$K\sim U$	$K\sim N$	$K\sim U$	$K\sim N$	$K\sim U$
Method — Table I										
Quadrature										
(1)	8.43-2	6.14-2	1.29-2	1.07-2	5.52-3	3.99-3	1.47-3	1.14-3	1.54-4	1.17-4
	$S = 2$		$S = 2$		$S = 3$		$S = 3$		$S = 4$	
(2)	5.56-2	4.01-2	1.95-2	1.39-2	7.42-3	5.57-3	1.42-3	1.50-3	5.90-4	6.02-4
	$S = 2$		$S = 2$		$S = 3$		$S = 5$		$S = 9$	
(4)	1.29-1	4.32-1	1.42-2	7.93-2	1.06-3	8.36-3	5.22-5	5.09-4	3.36-7	3.71-7
(6)	8.01-1	9.41-1	3.68-1	3.82-1	9.78-2	9.20-2	1.69-2	1.69-2	2.57-4	1.64-3

Explanation of the table is same as Table III except that $\rho = 0.75$.

computational costs $C = 81$ and $C = 400$ when the uniform weighting function, which gives equal weight to values of ϵ_t that are far from μ, is used.[19]

3.4 Changes in Specification of *W*

It is also worthwhile to establish some intuition about the robustness of the results in Table III to simple changes in the specification of W. One modification involved simply changing the constant in the W specification in Figure 1 from 13.71 to zero. As would be expected given the previous discussion, this change has no effect on most of the methods but leads to a substantial improvement in the non-normalized methods.[20] Changing W so that it is, on average, closer to zero leads to an improvement in this method because, as discussed earlier, for ϵ_t far from μ the non-normalized, self-interpolating Hermite method performs poorly because the weights w_i, $i = 1, \ldots, S$, are all very close to zero. The finding suggests that the normalized and non-normalized methods may not always be as different as indicated by Table III. Nonetheless, the use of the normalized methods for the interpolating approaches would seem to be generally advisable because they will often automatically avoid the problem of extremely bad approximation quality for values of ϵ_t far from μ.

The robustness of the results in Table III to changes in the first derivative of W was also examined by changing the coefficient on the ϵ_t term in Figure 1 from 3.60 to 18.60. This change has no effect on the interpolating methods but leads to a substantial decline in the quality of the normalized self-interpolating Hermite method.[21] Recall that the normalized self-interpolating Hermite method is a weighted average (in the sense that weights add to one) of actual W_{t+1} values. However, when ϵ_t is far below (above) μ, the values at which W_{t+1} are evaluated may all be substantially below (above) $E(\epsilon_{t+1} | \epsilon_t)$ until S becomes quite large. Thus, performance of the self-interpolating Hermite method relative to the interpolating methods worsens when W_{t+1} varies more with ϵ_{t+1}. Similarly, decreasing the coefficient on ϵ_t leads to an improvement in the performance of the self-interpolating Hermite method relative to the interpolating methods. Thus, the value of ρ above which the interpolating methods outperform the self-interpolating Hermite methods will depend on the shape of W_{t+1}.

4. EMPIRICAL EXAMPLE — TEACHER DECISIONS

In this section, the implementation of a dynamic, discrete choice model with a serially correlated state variable is described. The primary purpose of this section is to illustrate the feasibility of estimating such a model and to provide some informal evidence about the extent to which the approximation error in Section 3 influences the parameter estimates, predicted value functions, and estimated choice probabilities in which economists are ultimately interested. To this end,

[19] Although not shown, the simulation methods also perform substantially better when $\rho = 0.75$. For example, at $C = 81$ the average $E\hat{E}W_{t+1}$ value over the 100 simulation repetitions is 5.28-2, 6.72-2, 3.98, 4.70-1, 6.26, 5.75-1 for methods 7-12 in Table I respectively. Thus, the interpolating methods remain substantially better when $\rho = 0.75$.

[20] For example, at $C = 81$ the $E\hat{E}W_{t+1}$ value for the non-normalized, self-interpolating Hermite quadrature method decreased from 1.26 to 0.59 and the non-normalized self-interpolating Gauss–Legendre quadrature decreased from 3.12 to 0.74.

[21] For example, at $C = 81$ the $E\hat{E}W_{t+1}$ value for the normalized, self-interpolating Hermite quadrature method increased from 0.151 to 1.00. This is consistent with the theory in Rust (1997) which shows that the bounding constant for the worst-case error bounds involves the Lipschitz bound for the function in question. Increasing the derivative increases the Lipschitz bound. See Rust (1997, unpublished manuscipt) for a concrete illustration of this.

Table V. Descriptive statistics

Variable	Mean	Standard deviation
Number of years individual is observed (after certification)	9.0	4.1
Math SAT	476.2	93.2
Percentage female	72.5	
Number of children (in first year of certification)	0.2	0.5
Number of children (in 1986)	1.1	1.1
Percentage with at least one child (in first year of certification)	12.3	
Percentage with at least one child (in 1986)	63.2	
Percentage married (in first year of certification)	36.4	
Percentage married (in 1986)	77.7	
Percentage married in at least one period	81.4	
Number of years of post-bachelor education (as of 1986)	1.4	1.2
Years of teaching experience (as of 1986)	4.3	3.6
Years of non-teaching experience (as of 1986)	2.9	3.3

what is important is that the illustrative model be detailed enough to represent the type of issues that researchers encounter in practice. It is found that allowing serial correlation leads to more plausible estimates of certain model parameters. Nonetheless, it is not the intent of remainder of this paper to convince the reader that allowing serial correlation is of utmost importance in explaining behaviour in this particular application. Certainly there are other applications where modelling serial correlation is at the very essence of understanding behaviour.

The illustrative example involves a study of the occupational choices of elementary and secondary teachers after they become certified to teach. The data come from the National Longitudinal Study of the Class of 1972 and include between one and eleven years of information for each of 451 individuals who became certified to teach at some point between 1975 and 1985. Of the aggregated 4041 person years of data, 47.8% of years are spent teaching, 32.1% of years are spent working in non-teaching jobs, and the other 20.2% of the years are spent not working. Other basic descriptive statistics are shown in Table V. For more description of the data and a more in-depth discussion of sample selection, see Stinebrickner (2001a,b).

The specification of the model follows equations (1)–(6). In each period, the individual has either three or four choices (i.e. $|A(t)| = 3$ or $|A(t)| = 4$) depending on whether he or she is currently teaching. If the person is not teaching, he or she can choose either to remain out of the work force ($a = 1$), to accept a new non-teaching job offer ($a = 2$), or to accept a new teaching job offer ($a = 3$). If the person is currently teaching, he or she has the additional option of remaining in his or her current job ($a = 4$). It is assumed that the person receives the teaching and non-teaching job offers with probability one, but that there is randomness in the wage offer that is actually received. This will be discussed in more detail below.

For notational simplicity, let the vector X_t include both a set of observable characteristics of the individual that are assumed to be exogenous and known to the individual for all periods and a set of discrete variables that are endogenous but predetermined at time t given previous decisions. Variables of the first type include things such as sex and college entrance exam scores.[22] The state variables of the second type are the number of years of teaching and

[22] In reality, some characteristics included in X, such as Children and Marital status, are not truly exogenous and predetermined as assumed. This is discussed in more detail in Stinebrickner (2001b).

non-teaching work experience that the person has accumulated as of time t.

Let M^a_{it} and Q^{a}_{it} represent the wage and non-wage utility (in wage equivalents) respectively for person i at time t for some choice $a \in \{1, \ldots, |A(t)|\}$. The total current period reward (utility) that i receives in t by choosing a is assumed to be additive in M^a_{it} and Q^a_{it}. In a base model without unobserved heterogeneity, the current period reward is given by

$$u(\epsilon_{it}, a) = M^a_{it} + Q^a_{it} = \left[\alpha^a_M X_{it} + \epsilon_{it}(a)\right] + \left[\alpha^a_Q X_{it} + \nu_{it}(a)\right] \tag{16}$$

Thus, the vectors α^a_M and α^a_Q represent the effect that X_t has on the average wage and average non-wage utility received from option a. For the remainder of this section we suppress the person-specific subscript.

$\nu_t(a)$ represents randomness in the current period, non-wage utility of option a. $\epsilon_t(a)$ represents randomness in the wages associated with option a. If the person does not work, she is assumed to receive a wage of zero.[23] We make the simplifying assumption that the unobservable in the non-teaching wage equation, $\epsilon_t(2)$, is serially uncorrelated and is distributed $\mathrm{N}(0,\sigma^2_{\epsilon 2})$. We make the assumption that the unobservable in the wage from the new-teaching offer, $\epsilon_t(3)$, is also uncorrelated with previous wages and is distributed $\mathrm{N}(0,\sigma^2_{\epsilon 3})$. Serial correlation enters the model under the assumption that the unobservable in the wage equation for a particular teaching job follows the AR(1) process

$$\epsilon_t(4) = \rho\epsilon_{t-1}(a_{t-1}) + e_t$$

where $a(t-1) = 3$ or $a(t-1) = 4$ and $e_t \sim \mathrm{N}(0,\sigma^2_e)$.

Thus, the framework is identical to that described in Section 2. In particular, the choice-specific value functions are given by

$$v_t(\epsilon, a) = u(\epsilon, a) + \beta \int_{-\infty}^{\infty} W_{t+1}(\epsilon')p(\mathrm{d}\epsilon'|\epsilon) \quad \text{for } a = 1, \ldots, |A(t)| \tag{18}$$

As before, $W_{t+1}(\epsilon')$ is given by the right side of equation (5). As discussed in footnote 6, under the assumption that $\nu(1), \ldots, \nu(|A(\mathrm{t}+1)|)$ are iid extreme value, W_{t+1} has a closed-form solution

$$W_{t+1}(\epsilon') = \tau\left\{\lambda + \ln \sum_{a' \in A(t+1)} \exp[v_{t+1}(\epsilon', a')/\tau]\right\} \tag{19}$$

where λ is Euler's constant and $\tau^2\ \pi^2/6$ is the variance of the extreme value distribution. Note that, if the person accepts a teaching option at time t ($a = 3$ or $a = 4$), the person will consider three job offers in the next period so that $\epsilon_{t+1} = \{\epsilon_{t+1}(2),\ \epsilon_{t+1}(3)$, and $\epsilon_{t+1}(4)\}$ and the integral of interest in equation (18) is three-dimensional. Otherwise, the integral in equation (18) is two-dimensional.

The problems that arise when solving value functions are the same as those described in Section 2. These problems are addressed here by using the interpolating Hermite approximation method (row 1 of Table I) with a discretization approach that uses linear interpolation between the nearest grid points. However, in order to compute $\hat{W}_{t+1}$, we proceed in a slightly different

[23] That is, $\epsilon_t(a)$ and α^a_M are assumed to be zero for $a = 1$.

manner from that described earlier. In this application, $\epsilon_{t+1}(3)$ and $\epsilon_{t+1}(4)$ are each serially correlated with the wage that the person would receive if she stays in that job in the future. However, because the starting wage offers that arrive in the future for new jobs are assumed to be independent of current wages, $\epsilon_{t+1}(3)$ appears only in $v_{t+1}(\cdot,a=3)$ and $\epsilon_{t+1}(4)$ appears only in $v_{t+1}(\cdot,a=4)$.[24] As a result, rather than interpolating $W_{t+1}(\epsilon_{t+1}(2), \epsilon_{t+1}(3),\epsilon_{t+1}(4))$ directly, it is beneficial to interpolate $v_{t+1}(\epsilon_{t+1}(3), a=3)$ and $v_{t+1}(\epsilon_{t+1}(4), a=4))$ whenever necessary and to use these interpolated values along with values of $v_{t+1}(\cdot)$ and $v_2(\cdot)$ to approximate $W_{t+1}(\epsilon_{t+1}(2),\ \epsilon_{t+1}(3),\ \epsilon_{t+1}(4))$.[25] The benefit is simply that the interpolation of $v_{t+1}(\epsilon_{t+1}(4), a=4))$ and $v_{t+1}(\epsilon_{t+1}(3), a=3)$ takes place in one dimension in $\epsilon_{t+1}(4)$ and $\epsilon_{t+1}(3)$ respectively, whereas interpolating W_{t+1} directly would require interpolation in multiple dimensions.[26] Nonetheless, despite this slight difference, the issues are very similar to those described earlier.

Once value functions are solved by backwards recursion, estimation by maximum likelihood proceeds in a reasonably straightforward manner. In order to separately identify wages from non-pecuniary utility, it is necessary to use information on both the choices that individuals make and the wages that individuals receive. In particular, we are interested in the joint likelihood of all observed wages and all observed choices. If the wage offers associated with all jobs that a person considers are observed in all periods, the likelihood contribution for person i would be given by

$$L_i = P(\epsilon_1, \epsilon_2, \ldots, \epsilon_T)\ P(a(1), a(2), \ldots, a(T)|\epsilon_1, \epsilon_2, \ldots, \epsilon_T) \tag{20}$$

where $\epsilon_t = \{\epsilon_t(2),\epsilon_t(3),\epsilon_t(4)\}$ if $|A(t)| = 4$ and $\epsilon_t = \{\epsilon_t(2),\ \epsilon_t(3)\}$ if $|A(t)| = 3$, and $a(t)$ indicates the person's observed choice at time t. Thus, the first term is the joint density of all wages in all periods and the second term represents the joint probabilities of the choices made from years one to T conditional on the person's wages.

Unfortunately, wages are missing both because wage offers from jobs that are not accepted are not observed and because the survey design did not involve collecting wages in every year of a person's working career. Let $\epsilon = \{\epsilon_1, \ldots, \epsilon_T\}$ be the set of all wages in all years. Let $\epsilon^{\mathrm{O}} \subseteq \epsilon$ be the subset of all wages that are observed, and let $\epsilon^{\mathrm{U}} \subseteq \epsilon$ be the subset of all wages that are unobserved/missing. Note that $\epsilon^{\mathrm{O}} \cup \epsilon^{\mathrm{U}} = \epsilon$. The likelihood function with missing wages is found by integrating out the effects of the missing wages

$$L_i = \int P\big(a(1), a(2), \ldots, a(T)|\epsilon^{\mathrm{O}}, \epsilon^{\mathrm{U}}\big)P(\epsilon^{\mathrm{O}}, \epsilon^{\mathrm{U}})\mathrm{d}\epsilon^{\mathrm{U}} \tag{21}$$

which can be rewritten as

$$= \int P\big(a(1), a(2), \ldots, a(T)|\epsilon^{\mathrm{O}}, \epsilon^{\mathrm{U}}\big)P(\epsilon^{\mathrm{O}})P(\epsilon^{\mathrm{U}}|\epsilon^{\mathrm{O}})\mathrm{d}\epsilon^{\mathrm{U}} \tag{22}$$

Equation (22) can be simulated as

[24] The wage offer from a job that is not accepted does not influence either current period utility or the distribution of future wage offers.

[25] Note that computing W_{t+1} is not time consuming given $v_{t+1}(\cdot,a=1)$, $v_{t+1}(\cdot,a=2)$, $v_{t+1}(\cdot,a=3)$, and $v_{t+1}(\cdot,a=4)$ under the assumption that the ν's are iid extreme value.

[26] This implies that the computational cost of also allowing $\epsilon_{t+1}(2)$ to be serially correlated would be the cost of solving value functions for N $\epsilon_{t+1}(2)$ grid points in each period.

$$L_i^{\text{sim}} = \frac{1}{D}\sum_{j=1}^{D} P\big(a(1), a(2), \ldots, a(T)|\epsilon^{\text{O}}, \epsilon^{\text{U},j}\big)P(\epsilon^{\text{O}}) \tag{23}$$

where $\epsilon^{\text{U},j}$ is the jth of D draws from the density $P(\epsilon^{\text{U}}|\epsilon^{\text{O}})$.[27] Under the assumption that ϵ is normal, it is straightforward to compute the distribution of $(\epsilon^{\text{U}}|\epsilon^{\text{O}})$ and draw values of $\epsilon^{\text{U},j}$ from it. Given a particular draw of $\epsilon^{\text{U},j}$, the term $P(a(1),a(2),\ldots a(T)|\epsilon^{\text{O}},\epsilon^{\text{U},j})$ can be computed easily because, conditional on the wages, the year-specific choice probabilities are independent and the year-specific choice probabilities have a well-known closed form solution under the assumption that $\nu(1),\ldots,\nu(A)$ are iid extreme value and uncorrelated across time.[28] The wages in ϵ^{O} are not independent due to the serial correlation that exists within the wages at a particular teaching job. However, the joint density $P(\epsilon^{\text{O}})$ can be computed in a straightforward manner by writing it as the product of conditional densities.

Thus, the model can be estimated given a choice of S and a decision about how far apart to space the grid points. We refer to this spacing as Δ and note that the size of Δ should be viewed relative to the size of σ_{e3} which is estimated to be approximately 0.35. Given some basic intuition about the shape of W in $\epsilon(3)$ and $\epsilon(4)$, the earlier results suggest that the approximation error in $E\hat{W}_{t+1}$ should be quite small even if S is chosen to be relatively small and Δ to be relatively large. Nonetheless, the estimation of an actual dynamic programming model allows an informal examination of the extent to which approximation error in the $E\hat{W}_{t+1}$ values influences parameter estimates, predicted value functions, and estimated choice probabilities.

For the model described above, Figure 5 shows that the average parameter estimated using $S=6$ and $\Delta=0.1$ differs from its $S=5$, $\Delta=0.1$ counterpart by only 0.01 of its standard error. This average difference from the $S=6$ and $\Delta=0.1$ estimates increases in a non-linear fashion as S decreases and is 0.109 for $S=2$ and $\Delta=0.1$. Figure 6 shows that the average parameter estimated using $S=6$ and $\Delta=0.05$ differs from its $S=6$, $\Delta=0.1$ counterpart by only 0.013 of its standard error. This average difference from the $S=6$ and $\Delta=0.05$ estimates increases with Δ and is 0.511 for $S=6$ and $\Delta=0.9$. Thus, at $S=2$ and/or very large values of Δ, approximation quality is somewhat poor from the standpoint of the resulting parameter estimates. For example, it was found that at $S=2$ and $\Delta=0.9$ the average parameter estimate differs from its $S=6$, $\Delta=0.1$ counterpart by 0.487 of its standard error. However, it is important that the difference between parameter estimates obtained using $S=3,4,5$ with reasonable choices of Δ and parameter estimates obtained using $S=6$ and $\Delta=0.1$ is not particularly important relative to the sampling variation that is present. For example, at $S=3$ and $\Delta=0.4$, the average parameter estimate differs from its $S=6$, $\Delta=0.1$ counterpart by 0.096 of its standard error.[29]

[27] Equation (23) produces an unbiased estimate of the probability expression in equation (22). However, because $\text{E}\log(L_i^{\text{sim}}) \neq \log\text{E}(L_i^{\text{sim}})$, the simulated log likelihood function will only yield an unbiased estimated of the true log likelihood function as $D \to \infty$ (see, for example, Stern 1997). However, Borsch-Supan and Hajivassiliou (1993) show in some Monte Carlo experiments that, even for fixed D, simulated maximum likelihood parameter estimators have small asymptotic bias and root mean squared errors.

[28] Note that the closed-form solution for the choice probability at t will be a function of $v_t(\cdot,a=1)$, $v_t(\cdot,a=2)$, $v_t(\cdot,\epsilon_t(3),a=3)$, and $v_t(\cdot,\epsilon_t(4),a=4)$. Since the values of $\epsilon_t(3)$ and $\epsilon_t(4)$ associated with the observed wages will not typically be grid points, computing the likelihood function will itself require interpolation.

[29] It was found that a variable spacing method with an average spacing of 0.4 (obtained by reducing spacing for grid points that are closer to the unconditional mean) reduce this number from 0.096 to 0.051.

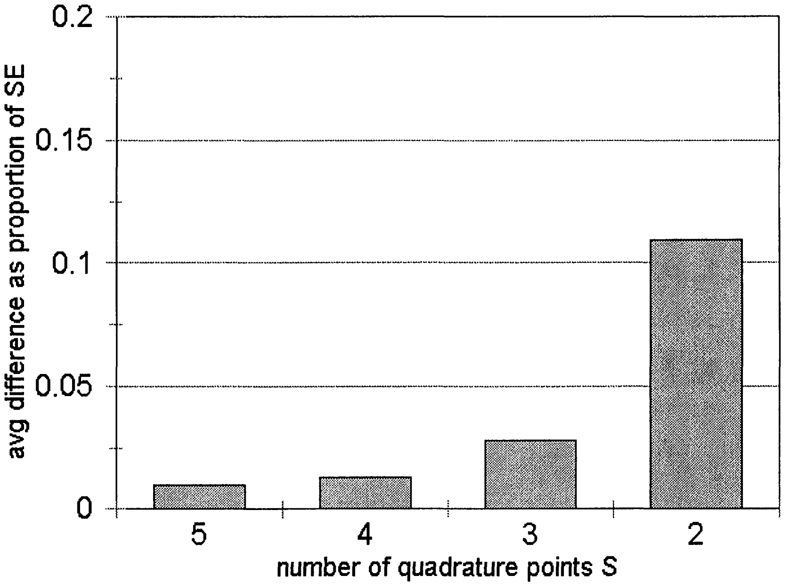


Figure 5. Parameter estimates for different S compared to $S = 6$ fixed delta

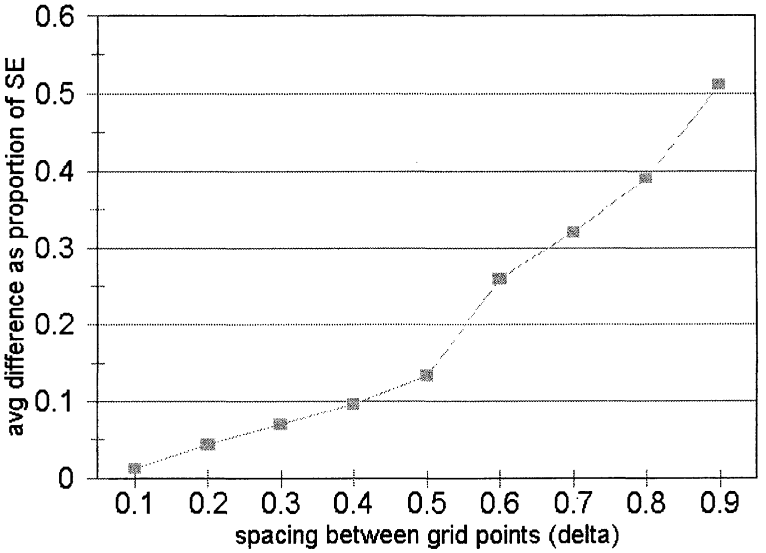


Figure 6. Parameter estimates different delta compared to delta = 0.05 fixed S. Note: the standard deviation of wage = 0.35, e.g. delta = 0.4 implies points are 1.14 s.d.'s apart

Table VI. Estimated value function for teaching $t = 4$ — different (S,Δ) combinations

	$\epsilon_3(3) = -2.4$	$\epsilon_3(3) = -1.6$	$\epsilon_3(3) = -0.8$	$\epsilon_3(3) = 0$	$\epsilon_3(3) = 0.8$	$\epsilon_3(3) = 1.6$	$\epsilon_3(3) = 2.4$
$S = 6, \Delta = 0.1$	11.620	12.425	13.259	14.330	16.554	20.369	24.997
$S = 3, \Delta = 0.4$	11.652	12.458	13.291	14.369	16.615	20.451	25.077
$S = 2, \Delta = 0.8$	12.169	12.974	13.805	14.884	17.085	20.774	25.273

For three S,Δ combinations, the table shows estimated $v_4(\epsilon(3), a = 3)$ for various values of $\epsilon(3)$.

Similar conclusions were reached with respect to the predicted value functions and estimated choice probabilities. For the $(S = 6,\ \Delta = 0.1)$, $(S = 3,\ \Delta = 0.4)$, and $(S = 2,\ \Delta = 0.8)$ combinations, Table VI shows how the estimated value functions $v_{t=4}(\epsilon(3), a = 3)$ vary with $\epsilon(3)$ for a person with 'median' characteristics. Across the seven values of ϵ_t in Table VI, the predicted value function for the $(S = 6,\ \Delta = 0.1)$ case differs (in absolute value) from the predicted value function for the $(S = 6,\ \Delta = 0.4)$ case by an average of 0.051 and differs from the predicted value function for the $(S = 2,\ \Delta = 0.8)$ case by 0.487. With respect to the estimated choice probabilities, the average absolute deviation (over all sample members and all choices) between the choice probabilities computed using $(S = 6,\ \Delta = 0.1)$ and the choice probabilities computed using $(S = 3,\ \Delta = 0.4)$ is 0.003. The average absolute deviation (over all sample members and all choices) between the choice probabilities computed using $(S = 6,\ \Delta = 0.1)$ and the choice probabilities computed using $(S = 2,\ \Delta = 0.8)$ is 0.015.

As shown by comparing the results for the $(S = 2,\ \Delta = 0.8)$ case and the $(S = 6,\ \Delta = 0.1)$ case, poor approximation quality in computing $\hat{EW}_{t+1}(\epsilon_t)$ can translate into non-trivial differences in predicted value functions and estimated choice probabilities. However, the fact that the $(S = 3,\ \Delta = 0.4)$ case leads to 'similar' parameter estimates, predicted value functions, and estimated choice probabilities as the $(S = 6,\ \Delta = 0.1)$ case is important because the estimation using the latter takes approximately 33 times as long as estimation using the former. In this application, this time savings is crucial in order to estimate a more elaborate specification of the model in which the non-pecuniary utility that a person receives from a particular option, a, depends on a person-specific heterogeneity component.[30]

The estimates of the full model with heterogeneity are shown in Table VII. The first column shows the model estimates under the assumption that $\rho = 0$. The second column shows the model estimates when ρ is estimated. The estimate for ρ, 0.923, is quantitatively large and statistically significant. The loglikelihood value is substantially better, -4686 versus -4233, when serial correlation is allowed.[31] Not surprisingly, the biggest differences between the $\rho = 0$ and $\rho \neq 0$ cases are found in the estimated effects of the time trend variables and the teaching experience variable in the teaching wage equation. With respect to the former, it is well known that a U-shaped time trend was present in teachers' wages during the 1975-1985 period.[32] The

[30] Letting the heterogeneity term for option a be denoted $\eta_i^a, u(\epsilon_{it}, a) = M_{it}^a + Q_{it}^a = \left[\alpha_M^a X_{it} + \epsilon_{it}(a)\right] + \left[\alpha_Q^a X_{it} + \eta_i^a + \nu_{it}(a)\right].$ The assumption that $\eta_i^3 = \eta_i^4$ is made.

[31] From a substantive standpoint, it is important to note that the estimated specifications do not examine the degree to which the serial correlation in wages would be reduced if a permanent unobserved heterogeneity term was included in the wage equation for teachers.

[32] See for example, US Department of Education, National Center for Education Statistics, *The Condition of Education*, Washington, DC. 1994.

Table VII. Structural model estimates

Variable	Specification 1 $\rho = 0$	Specification 2
Teaching wage		
CONSTANT	5.405* (0.069)	5.718* (0.046)
TIME (1975 = 1, 1976 = 2, . . .)	−0.151* (0.017)	−0.270* (0.013)
*TIME*TIME*	0.008* (0.001)	0.015* (0.001)
MALE	−0.015 (0.019)	0.007 (0.012)
SAT	−0.020* (0.010)	−0.009 (0.007)
Years of post-bachelor education	0.046* (0.009)	0.053* (0.009)
Years of teaching experience	0.050* (0.006)	0.019* (0.005)
Years of non teaching experience	−0.001 (0.010)	0.006 (0.009)
Autoregressive coefficient		
ρ	•	0.923* (0.008)
Non-teaching wage		
CONSTANT	4.509* (0.092)	4.517* (0.094)
TIME (1975 = 1,1976 = 2, . . .)	−0.011 (0.026)	−0.015 (0.027)
*TIME*TIME*	0.003 (0.002)	0.002 (0.002)
MALE	0.077* (0.029)	0.078* (0.029)
SAT	0.025* (0.012)	0.024* (0.012)
Years of post-bachelor education	0.039* (0.010)	0.039* (0.008)
Years of teaching experience	−0.039* (0.009)	−0.36* (0.010)
Years of non teaching experience	0.068* (0.009)	0.061* (0.009)
Teaching non-pecuniary utility		
CONSTANT	−3.445* (0.157)	−3.384* (0.175)
MALE	0.039* (0.102)	0.078* (0.115)
SAT	−0.018 (0.030)	−0.032 (0.031)
Years of teaching experience	−0.085* (0.007)	−0.081* (0.007)
Years of non-teaching experience	−0.057* (0.011)	−0.065* (0.009)
Number of children — *CHILD*	−0.438* (0.025)	−0.468* (0.031)
CHILD × *MALE*	0.432* (0.049)	0.452* (0.062)
MARRIED	−0.428* (0.063)	−0.426* (0.070)
MARRIED × *MALE*	0.371* (0.095)	0.350* (0.115)
Non-teaching non-pecuniary utility		
CONSTANT	−3.757* (0.157)	−3.860* (0.173)
MALE	−0.114 (0.105)	−0.072 (0.116)
SAT	−0.027 (0.030)	−0.025 (0.032)
Years of teaching experience	−0.049* (0.009)	−0.060* (0.010)
Years of non-teaching experience	−0.044* (0.009)	−0.035* (0.009)
CHILD	−0.408* (0.026)	−0.418* (0.032)
CHILD × *MALE*	0.396* (0.049)	0.408* (0.062)
MARRIED	−0.463* (0.063)	−0.448* (0.070)
MARRIED × *MALE*	0.471* (0.097)	0.445* (0.112)
Variance terms		
$\eta_3 = \eta_4$ (heterogeneity teaching)	0.072* (0.016)	0.086* (0.017)
η_2 (heterogeneity non-teaching)	0.083* (0.017)	0.132* (0.019)
η_3 (heterogeneity home)	0.920* (0.041)	0.909* (0.053)
$\sigma_{\epsilon 3}$	0.371* (0.005)	0.352* (0.004)
σ_{e}	0.384* (0.006)	0.310* (0.004)
$\sigma_{\epsilon 2}$	0.460* (0.007)	0.453* (0.006)
τ	0.417 (0.018)	0.441* (0.017)
Log likelihood function	−4686.11	−4233.30

Note: The numbers are estimates from a specification in which the discount factor, β, is set to 0.95. The numbers in parentheses are asymptotic standard errors. * denotes an asymptotic t ratio greater than two. Specification 1 is the model which does not accommodate serially correlated wage error terms ($\rho = 0$). Specification 2 is the model which allows serial correlation ($\rho \neq 0$).

$\rho \neq 0$ estimates imply a more pronounced U-shape than the $\rho = 0$ estimates. With respect to the latter, in the $\rho = 0$ case, an additional year of experience is estimated to increase a teacher's wage by approximately 5%. In the $\rho \neq 0$ case, an additional year of experience is estimated to increase a teacher's wage by approximately 2%. The smaller estimate seems more plausible given previous literature on the wages of teachers. For example, using data from the High School and Beyond Survey, Lee and Smith (1990) ,estimate that an additional year of teaching experience increases teaching wages by slightly less than 3%. Using their own survey data, the research department of the American Federation of Teachers (1993) calculates that 'the typical state awards an average of ... 2.4 percent of the average salary for each year between the starting salary and the average salary'.[33]

The estimated effects of the variables in the non-teaching wage equation and the non-pecuniary equations do not tend to change substantially between the $\rho = 0$ case and the $\rho \neq 0$ case. In both cases, the non-pecuniary estimates indicate that family variables play a very important role in determining whether a woman remains in the workforce. Women with children and married women receive significantly less non-pecuniary utility from working in either a teaching job or a non-teaching job (relative to being out of the work force), and, as a result, are more likely than other women to be out of the work force. The family variables have very little effect on men.

5. CONCLUSIONS

Testing in Section 3 suggests that the relative approximation quality of the various methods for computing $E\hat{W}_{t+1}(\epsilon_t)$ may vary substantially with a researcher's particular application. Given the normality assumption for the transition distribution, the interpolating method based on Hermite quadrature, the interpolating method based on Gauss–Legendre quadrature, and the self-interpolating method based on Hermite quadrature were found to be the most promising methods. The interpolating methods hold a large advantage when the degree of serial correlation, as measured by ρ, is high. As discussed, this finding arises because the self-interpolating Hermite methods perform very poorly when ϵ_t is far from the mean of the ϵ process. However, for smaller ρ, the self-interpolating Hermite methods perform very well for all values of ϵ_t that are reasonably likely to arise. For the base specification of W_{t+1}, the threshold below which the self-interpolating Hermite outperforms the interpolating methods was found to be approximately 0.80 when the normal weighting function was used for $K(\epsilon_t)$. However, as demonstrated, this number would tend to be lower (higher) under a specification in which W_{t+1} varies more (less) with ϵ_{t+1}. The findings also indicate that the use of a normalized weighting scheme with the self-interpolating methods is generally advisable and that, for a fixed computational cost C, the optimal choice of the number of quadrature points for the interpolating methods will typically be substantially smaller than the number of quadrature points associated with the self-interpolating methods (i.e. smaller than $\sqrt{C}$).

Careful consideration of the choice of approximation method seems important because, as suggested by the empirical example in Section 4, poor approximation quality of $E\hat{W}_{t+1}$ can have a non-trivial effect on estimated parameters, predicted value functions, and estimated choice

[33] See *Survey and Analysis of Salary Trends*, Research and Information Services Department, American Federation of Teachers, AFL-CIO (1993).

probabilities. Nonetheless, from the standpoint of the future estimation of realistic dynamic, discrete choice models, it is promising that the estimated parameters, predicted value functions, and estimated choice probabilities in the empirical example appear to be quite accurate for very feasible choices of the number of quadrature points and the spacing between grid points.

APPENDIX: PARAMETRIC AND DISCRETIZATION APPROACH FOR COMPUTING $\hat{W}_{t+1}(\epsilon_{t+1})$

The interpolating methods described in Section 2 require the approximation of the $\hat{W}_{t+1}(x_i(\epsilon))$ values that appear in equation (7). In Section 2, a connection was made between the 'discretization' and 'parametric' approaches for computing $\hat{W}_{t+1}(\epsilon_{t+1})$ and the 'parametric' approach when θ in equation (10) is estimated using OLS. Here, we examine the relationship using the testing framework discussed in Section 3. We concentrate on the interpolating Hermite quadrature case, and calculate our estimate of EEW_{t+1} using the normal K. As before, we use W as in Figure 1 and $\rho = 0.95$. We isolate the difference between the discretization and parametric method by examining the $S = 7$ case so that any approximation error that is present in $E\hat{W}_{t+1}(\epsilon_t)$ for any ϵ_t arises entirely from the approximation error in the $\hat{W}_{t+1}(\epsilon_{t+1})$ values. For the discretization approach, $E\hat{W}_{t+1}(\epsilon_t)$ is computed, as before, using linear interpolation between equally spaced grid points. For the parametric approach, attention is limited to polynomial specifications for the parametric function $W_{t+1}(\epsilon,\theta)$ so that $\hat{W}_{t+1}(\epsilon) = \sum_{k=1}^{K} \hat{\theta}_k \epsilon^{k-1}$.

Because W is a polynomial of degree 13 in ϵ_t, the interpolating Hermite quadrature approximation is exact for $S = 7$ if $\hat{W}_{t+1}(\epsilon_{t+1})$ is exact for $W_{t+1}(\epsilon_{t+1})\forall\epsilon_{t+1}$. Under the parametric approach, $\hat{W}_{t+1} = W_{t+1}\ \forall\epsilon_{t+1}$ if $K = 14$. Thus, EEW_{t+1} is zero in this case. Table AI concentrates on the non-trivial case in which $K \leq 13$. Obtaining insight into the relationship between misspecification of the parametric function used to compute $\hat{W}_{t+1}(\epsilon_{t+1})$ and approximation error in $E\hat{W}_{t+1}(\epsilon_t)$ is of interest. Although the Weirstrauss theorem indicates

Table AI. Interpolating Hermite quadrature — discretization and parametric approximation of W_{t+1}

N	15	30	50	100	1000
Discretization — local interpolating	2.87-2	6.88-3	2.42-3	5.92-4	5.89-6
Parametric, $K = 2$	5.96-1	5.81-1	5.75-1	5.71-1	5.66-1
Parametric, $K = 4$	4.10-1	4.01-1	3.97-1	3.94-1	3.91-1
Parametric, $K = 6$	2.66-1	2.58-1	2.53-1	2.50-1	2.46-1
Parametric, $K = 8$	1.26-1	1.28-1	1.24-1	1.21-1	1.17-1
Parametric, $K = 10$	2.27-2	3.42-2	3.43-2	3.29-2	3.08-2
Parametric, $K = 12$	6.36-4	3.09-3	3.45-3	3.37-3	3.07-3
Parametric, $K = 13$	6.21-4	2.87-3	3.22-3	3.15-3	2.88-3

Note: The table shows $EE\hat{W}_{t+1}$ from equation (15) for interpolating Hermite quadrature methods. W is as in Figure 1, $\sigma_e = 0.3$, and $\rho = 0.95$. The first row involves discretization method for $\hat{W}_{t+1}$ with linear interpolation between N grid points that are evenly spaced over the interval described in footnote 17. Additional rows show parametric method for $\hat{W}_{t+1}$ with the parametric approximation function

$$\hat{W}_{t+1}(\epsilon) = \sum_{k=1}^{K} \hat{\theta}_k \epsilon^{k-1}$$

For the parametric cases, N represents the number of points that are used in the regression that determines $\hat{\theta}_1, \ldots, \hat{\theta}_K$. The points are equally spaced over the interval described in footnote 17.

that any function can be approximated arbitrarily well (on a closed and bounded interval) by a polynomial of high enough degree, in practice W will not be a polynomial and misspecification will occur given finite K.

As expected, the last column of Table AI shows that the amount of approximation error (as measured by $E\hat{E}W_{t+1}$) increases as the interpolating function becomes 'more' misspecified (i.e. for smaller K). Looking across any particular row in Table AI indicates that increasing the number of grid points N does not generally have a substantial effect on approximation quality if the interpolating function is misspecified.[34] Table AI makes it clear that, beyond a certain N value, substantial improvements in the approximation quality for the parametric approach can only be obtained by increasing the degree of flexibility in the polynomial function, not by increasing N.[35]

Table AI generally indicates that the specification of the parametric $W_{t+1}(\epsilon,\theta)$ is important. Unfortunately, it is often not possible to know much about the shape of W before the model is estimated (and estimation requires a specification of the parametric approximation function). While one can protect against approximation error by increasing the flexibility of the parametric approximation function (e.g. in this case by increasing K), doing so increases the number of interpolating points that are necessary to identify the coefficients of the interpolating function and creates more coefficients to estimate. These problems may not be significant in one-dimensional cases but may become more worrisome when attempting to specify the parametric approximation function for a multi-dimensional case.

Under the discretization approach with local linear interpolation, $\hat{W}_{t+1}(\epsilon_{t+1}) \to W_{t+1}(\epsilon_{t+1})$ as $N \to \infty$.[36] As a result, it should be the case that, whenever the parametric approximation function is misspecified, there will exist a N^* such that the methods with discretization outperform the methods with parametric approximation whenever $N > N^*$. In this case, Table AI shows that if the parametric function is specified with $K \leq 8$, N^* is less than 15. For $\mathrm{K} = 10$, N^* is slightly greater than 15. For $K = 12$ and $K = 13$, N^* is between 30 and 50.

ACKNOWLEDGEMENTS

The author would like to thank Steven Stern, John Rust, Jaap Abbring, Chris Ferrall, Michael Brien, Chris Swann, anonymous referees, and seminar participants for useful comments and suggestions at various stages of this work. The work was made possible in part by support from the National Institute of Aging (r01-AG17579). The author also appreciates the warm hospitality he received at the Osceola Research Institute.

[34] Note that although Keane and Wolpin (1994) discuss the possibility of using their method to allow for serial correlation, the emphasisof their paper does not focus on the serial correlation case and the majority of their tests involve the use of state variables that are discrete. When state variables are discrete, the grid points can be chosen to correspond to actual elements of the state space. As they discuss, this implies that increasing the number of grid points necessarily decreases the number of values of ϵ_{t+1} for which W_{t+1} must be approximated and convergence of $E\hat{W}_{t+1}(\epsilon_{t+1})$ to the truth can be obtained by allowing N to approach the true number of state points. This convergence criterion is not particularly relevant in the context of a model with serial correlation because the 'discretized' number of possible values of ϵ_{t+1} will typically be extremely large (in the application of Section 4, the number is in the millions for many periods).

[35] Of course, it is important to note that the increasing N beyond small levels may tend to be more advantageous when the approximation/interpolation is taking place in higher dimensions.

[36] Under this method, $\hat{W}_{t+1}(\epsilon_{t+1})$ is computed as a linear combination of the value functions $W_{t+1}(\epsilon^{i*}_{t+1})$ and $W_{t+1}(\epsilon^{j*}_{t+1})$ for the nearest, surrounding grid points ϵ^{i*}_{t+1} and ϵ^{j*}_{t+1}. As N becomes large, ϵ^{i*}_{t+1} and ϵ^{j*}_{t+1} become close to ϵ_{t+1}.

REFERENCES

Bellman, R. (1957), *Dynamic Programming*, Princeton University Press, Princeton, NJ.

Berkovec, J. and S. Stern (1991), 'Job exit behavior of older men', *Econometrica*, **59**, 189–210.

Borsch-Supan, A. and V. Hajivassiliou (1993), 'Smooth unbiased multivariate probability simulators for maximum likelihood estimation of limited dependent variable models', *Journal of Econometrics*, **58**, 3477–368.

Christenson, B. (1990), Estimation of dynamic programming models. PhD Dissertation, Cornell University.

Härdle, W. and O. Linton (1994), 'Applied nonparametric methods', R. F. Engle and D. L. McFadden (Eds), *Handbook of Econometrics*, North Holland: New York Vol. IV, Chapter 38.

Heckman, J. and B. Singer (1984), 'A method for minimizing the impact of distributional assumptions in econometric models', *Econometrica*, **52**, 360–399.

Hubbard, G., J. Skinner and S. Zeldes (1995), 'Precautionary saving and social insurance', *Journal of Political Economy*, **103**.

Judd, K. L. (1998), *Numerical Methods in Economics*, MIT Press, Cambridge MA.

Keane, M. P. and K. I. Wolpin KI (1994), 'The solution and estimation of discrete choice dynamic programming models by simulation and interpolation: Monte Carlo evidence', *Review of Economics and Statistics*, **76**, 648–672.

Keane, M. and K. Wolpin (1997), 'The career decisions of young men', *Journal of Political Economy*, **105**, 473–522.

Lee, V. and J. Smith (1990), 'Gender equity in teachers' salaries: a multilevel approach', *Educational Evaluation and Policy Analysis*, **12**, 57–81.

Miller, R. (1984), 'Job matching and occupational choice', *Journal of Political Economy*, **92**, 1086–1120.

Pakes, A. (1986), 'Patents as options: estimates of the value of holding European patent stocks', *Econometrica*, **54**, 755–784.

Rust, J. (1987), 'Optimal replacement of GMC bus engines: an empirical model of Harold Zurcher', *Econometrica*, **55**, 999–1033.

Rust, J. (1997), 'Using randomization to break the curse of dimensionality', *Econometrica*, **65**, 781–832.

Rust, J. and C. Phelan (1997), 'How social security and medicare affect retirement behavior in a world of incomplete markets', *Econometrica*, **65**, 487–516.

Stern, S. (1997), 'Simulation based estimation', *Journal of Economic Literature*, **35**, 2006–39.

Stinebrickner, T. (1996), *A Dynamic, Discrete Choice Model of Teacher Attrition*, PhD dissertation, University of Virginia.

Stinebrickner, T. (1998), 'An empirical investigation of teacher attrition', *Economics of Education Review*, **17**, 127–136.

Stinebrickner, T. (1999), 'Estimation of a duration model in the presence of missing data', *Review of Economics and Statistics*, 529–542.

Stinebrickner, T. (2001a), 'A dynamic model of teacher labor supply', *The Journal of Labor Economics* (forthcoming).

Stinebrickner, T. (2001b), 'Compensation policies and teacher decisions', *The International Economic Review* (forthcoming).

Stroud, A. H. and D. Secrest (1966), *Gaussian Quadrature Formulas*, Prentice Hall, Englewood Cliffs, NJ.

Tauchen, G. and R. Hussey (1991), 'Quadrature-based methods for obtaining approximate solutions to nonlinear asset pricing models', *Econometrica*, **59**; 371–396.