School of Economic Sciences

Working Paper Series
WP 2011-7

Sequential Self-Selection of Program Adherence

By

Ron Mittelhammer, Robert Rosenman, and Vidhura Tennekoon

September 2011

WASHINGTON STATE UNIVERSITY
World Class. Face to Face.

Sequential Self-Selection of Program Adherence

Ron Mittelhammer, Robert Rosenman, and Vidhura Tennekoon*

Abstract

Conceptual modeling and empirical analysis of individuals' sequential self-selection of adherence to voluntary treatment or education programs is an ongoing and unsettled area of inquiry. Both the representation of the decision process and the implementation of econometric methods for estimating the unknowns of such models are difficult, especially when the decision process allows repeated exit and reentry to the program. We present and apply a conceptual model that is both consistent with the random utility model structure and capable of representing complex patterns of self-selected adherence while being relatively straightforward to specify and implement empirically. The approach is based on a "distributed error" form of stochastic heterogeneity that exploits the use of singular normal distributions and leads to efficiency in the representation and computation of maximum likelihood estimates of model parameters. The model can be straightforwardly applied to non-normal distributions as well, and to more elaborate nonlinear specifications of the systematic drivers of self-selected adherence decisions.

JEL Codes: C35, C50, C5, I10

Key Words: sequential selection, random utility model, program adherence, compliance, distributed error stochastic heterogeneity

---

* Respectively, Regents Professor of Economics and Statistics, Professor of Economics, and PhD candidate in the School of Economic Sciences, Washington State University, Pullman, WA, 99164-6210. The study was supported in part by the National Institute of Drug Abuse (grants R21 DA025139-01Al and R21 DA19758-01). We thank the program providers and families who participated in the program evaluation, and the NIH Grant Working Group for their comments and support.

1.0 INTRODUCTION

Curative, preventive, training, and educational programs often entail repeated commitments by participants. By this we mean that the program requires multiple, sequential adherence decisions by participants regarding prescribed actions or behaviors, including taking medication, attending sessions or classes, or doing work on one's own in some prescribed manner. No matter how efficacious a program, if participant fidelity through the sequence of prescribed commitments is lacking, the effectiveness of the program will almost invariably be negatively affected. In fact, nonadherance is the fundamental difference between what is known as the "complier average causal effect" (the effect of the treatment for full adherence) and the "intent to treat" effect (the average effect of the treatment on those assigned to the treatment group) (Stuart, Perry, Le and Ialongo, 2008).

If participation in the elements of a program is voluntary and requires sequential commitments, the eventual outcome for any participant depends on the actual choices made regarding adherence to the prescribed program. The pattern of choices self-selects the administered input from a program, which can include such features as dosage of medication, intensity of intervention, the specific types of treatment actually received, or other characteristic features of any program requiring sequential adherence to program elements.

Sequential adherence issues affect treatments in a wide variety of programs, such as schooling (Angrist and Imbens, 1995), breast cancer treatment (Basu, Heckman, Navaro-Lozano and Urzua, 2007), mental health (Chib and Jacobi, 2008), mastitis treatment in cows (Crouchley and Ganjali, 2002) and substance abuse (Lu and McGuire, 2002). One of the most studied areas of participant sequential self-selection regards adherence by patients to prescribed medication, mostly because adherence is so poor (Sackett, 1976; McDonald, Garg and Haynes, 2002), increasing health care costs by about $100 billion a year (Osterberg and Blaschke, 2005) while seriously compromising economic evaluations (Rosen, et al., 2009).

In this paper we offer a relatively straightforward, parsimonious, and readily implementable empirical model of a complex decision problem – that of selecting the pattern of adherence to a sequence of treatments within a prescribed program of treatments. The basic model we present provides a consistent behavioral context for rationalizing observed patterns of treatment adherence that has at its foundation a random utility model that accounts for observable factors influencing choices as well as accumulated learning derived from experiencing past treatments. While parsimonious, the model is rich enough so that results obtained from its application can provide policy makers and program providers with an analytical way of rationalizing, anticipating, and affecting treatment adherence behavior. We illustrate the model's relative ease of application, its sampling properties, and its interpretation through a combination of Monte Carlo simulations and a substantive empirical application which analyzes adherence to a nationally recognized parenting and family strengthening program.

In the following section we review literature that is particularly relevant for empirically modeling sequential program adherence decisions. In section 3 we provide a general conceptual approach for modeling sequential adherence decision processes. We propose an empirical specification and methods for estimating the model in section 4, and implement the methods in a Monte Carlo experiment to illustrate the computational and repeated sampling behavior of the approach. We concentrate on a detailed presentation of the more parsimonious "rank one" rendition of the model in this paper. However, we also indicate how the model can be generalized in various ways to accommodate more complex systematic and stochastic representations of decision maker (DM) behavior. Then, in section 5, we apply the model to data relating to a family strengthening program designed to reduce substance abuse that consists of multiple educational sessions. We conclude with implications, both from the point of view of econometric modeling in sequential adherence settings and substantive findings from the empirical application, and suggest directions for future research.

2.0 EMPIRICAL MODELS OF SEQUENTIAL PROGRAM ADHERENCE

How should one model and empirically estimate sequential program adherence?[1] One possibility is to model the selection problem as a polychotomous choice situation, where individuals choose one option from among all potential combinations of sequential program event patterns leading to a multinomial decision model, such as a multinomial logit or probit model. A key consideration in this approach is the extent to which the polychotomous choice models impose underlying structure beyond index restrictions. One example of what we mean is presented in Trost and Lee (1984), who use a variant of McFadden's (1974) conditional logit model in a random utility characterization of schooling choice, which results in a polychotomous model. As is common in such applications, they assume errors in their random utility function are independently and identically Gumbel distributed, an assumption of convenience that is necessary to rationalize the logit approach but is not necessarily appropriate for a given sequential decision problem. Lee (1995) later developed a

[1] We emphasize that we are interested in *patterns* of choice, and not only variable treatment intensity, which was the main focus of Angrist and Imbens (JASA,1995)

semiparametric maximum likelihood method for estimating polychotomous choice methods that can be extended to sequential choice models if the pattern to be followed is made "once and forever" at the beginning, thus, rendering it inappropriate if learning is part of the process.

Despite the large literature on polychotomous choice, including that directed towards treatment selection, there have been few instances where the underlying process leading to adherence decisions has been explicitly modeled. Barrios (2004) offers a general approach to incorporating selection through a random utility model. However, his approach is not applicable in the current decision context because the adherence decisions in his model are not sequential. Three leading examples of papers which do model sequential decisions are Diggle and Kenward (1994), Lahiri and Song (2000) and Heckman and Navarro (2007).

Diggle and Kenwood explain the dropout mechanism for repeated responses, which allows for dependence between the likelihood of dropping out and past experience. Heckman and Navarro discuss semiparametric identification in a similar stopping model for schooling. Like Diggle and Kenwood they specifically disallow returning to school once a student drops out, noting (in their footnote 45) that a useful generalization would allow students to return to schooling "at different times as information sets are revised."[2] Lahiri and Song also disallow a return to previous behavior in a dynamic model of smoking. Their model of sequential choices has individuals choose to be a smoker or a nonsmoker in the first period, and then in a second period has them choose into one of three groups; nonsmokers, ex-smokers, and current smokers. The result is a switching regression model with sequential self-selection. However,

[2] This also explains why Lee's (1995) approach is not applicable to treatment adherence problems, since it does not allow reentry after dropout based on *revised* information. The structure of the errors implies that the pattern is chosen once and for all as treatment commences because it does not explicitly allow updating through time.

like Diggle and Kenwood and Heckman and Navarro, they assume once out, always out – that is, once a person becomes a nonsmoker, she is always a nonsmoker.[3]

The general model structure that we propose provides a method of generalizing the schooling problem to allow for repeated reentries, as Heckman and Navarro allude to in their footnote. We model a sequence of discrete adherence decisions which together comprise a prescribed program. At each step in the sequence the decision maker must decide if she will accept the next stage of the prescribed program, which we henceforth refer to generically as a "treatment"[4]. However, unlike the three papers just discussed, skipping or missing a treatment does not preclude the possibility of taking future treatments, and in any combination. Based on sequentially updated information relating to a DM's experience with the treatments, a decision about subsequent treatments is made. This process continues through the completion of program treatment decisions.

3.0 A GENERAL STRUCTURE FOR THE SEQUENTIAL ADHERENCE DECISION PROCESS

Consider a DM contemplating a sequential course of J treatments. At each treatment decision point, j, she must choose whether or not to take the treatment, with the anticipated utility of doing so being conditioned on the sequence of treatment experiences accumulated prior to decision point j. The ultimate observed pattern of treatments received is the result of a sequentially accumulated set of experiences that update the anticipated utilities on which

[3] In addition, they do not allow nonsmokers to move into the smoking ranks.

[4] Although we choose for convenience to adopt the single word "treatment", we underscore that our analysis applies generally to a wide variety of interpretations of "treatments", including, for example, drug and medical therapies, employment training and other more general educational programs, psychological and other therapy experiences, and generic skill-building programs, such as in settings relating to the empirical application presented later in the paper.

decisions to receive treatments are based. There is the possibility that an experiential threshold is reached whereby it is perceived that insufficient benefit would be gained from any future treatments, at which point a decision to stop a program of treatment is made and no further treatments are received (i.e., the "drop out" problem). Alternatively, the DM might decide that certain remaining treatments appear worthwhile, but others are not, thus selectively choosing to forego only a subset of the remaining treatments, and returning after missing one or more treatments. After each treatment, the DM evaluates her future participation anew.

Given the preceding structure for making decisions, it is clear that a model of the treatment decision process cannot be one that proceeds as if the decision to receive treatments in a particular pattern is made *ex ante*, once and for all, prior to the commencement of the program. Rather, the decision process is a sequence of decisions made over time, and decisions about receiving future treatments are informed through "learning-by-doing" and considering the associated accumulated prior experiences relating to all treatments received to that point.

Let $\mathbf{U}_i^A\left(\mathbf{X}_{it} \mid \mathbf{D}_{it}\right)$ represent a $J \times 1$ state vector of anticipated utilities for DM_i, who is at the point of making her decision regarding the t^{th} treatment in a course of J treatments[5]. The $K \times 1$ column vector $\mathbf{X}_{it}$ represents the state of relevant sociodemographic and treatment characteristics existent at decision point *t*, which frame the decision environment of DM_i at the

[5] We follow the convention of denoting vectors and matrices in bold font, and scalars, other than numbers, in non-bold italics.

decision point[6]. The $J \times 1$ vector, $\mathbf{D}_{it}$, of binary indicators represents the cumulative historical state of DM_i's treatment acceptances (1's) or rejections (0's) as of decision point *t*, which defines the pattern of program treatments the DM has actually experienced to date.

At the initial decision point, $t=1$, $\mathbf{D}_{i1}=[0,0,...,0]'$, i.e., no treatments have been experienced. The j^{th} entry in the utility vector, $U_{ij}^{A}(\mathbf{X}_{it} \mid \mathbf{D}_{it})$, is either the current anticipated utility of future treatment j if $j>t$, which can be updated by future treatment experiences, or else the *final anticipated utility* on which an acceptance-rejection decision was based if $j \le t$. By allowing updating of future anticipated utilities based on treatment experience, it is explicitly assumed that there are elements to the treatment the DM cannot predict and, upon experiencing the treatment, she uses the realization of those elements to update future anticipated utilities. Assuming that the utility index is denominated such that it represents the net utility of receiving the j^{th} treatment compared to the next best use of time and expense involved in receiving the treatment, the DM will decide to take treatment j if the anticipated utility for the treatment at decision point $j=t$ is positive, that is, if $U_{ij}^{A}(\mathbf{X}_{ij} \mid \mathbf{D}_{ij})>0$, and will decide not to take the treatment otherwise.

The decision process proceeds sequentially, through the J decision stages, based on successive updates of $\mathbf{U}_{ij}^{A}(\mathbf{X}_{it} \mid \mathbf{D}_{it})$, for $j \ge t=1,...,J$, resulting from decisions made and

[6] Characteristics might change in some treatment programs. For example, one possible explanatory variable is physical vigor. A course of radiation or chemo therapy might debilitate a cancer patient so that the level of physical vigor changes. The model accommodates time dependent characteristics, including those that reflect observed systematic updates to factors that affect the decision environment in a material way, by including such factors in the specification of $\mathbf{X}_{it}$, recognizing that the vector can change with *t*.

experiences realized from treatments. The process begins with an initial decision to participate in the program of treatments, where the DM will forgo the entire treatment program *iff* $\mathbf{U}_i^A(\mathbf{X}_{i1} \mid \mathbf{D}_{i1}) \leq [\mathbf{0}]$. The DM will accept at least some of the treatments if $U_{ij}^A(\mathbf{X}_{i1} \mid \mathbf{D}_{i1}) > 0$ for some j. The DM will have taken all of the prescribed set of treatments in the treatment program if $U_{ij}^A(\mathbf{X}_{ij} \mid \mathbf{D}_{ij}) > 0 \ \forall j$, in which case $\mathbf{D}_{it_{if}} = [1,1,...,1]'$, where t_{if} denotes the state following the final program adherence decision made by DM_i, i.e., the state at which the treatment program is rendered complete by the DM's decisions. The DM will quit the program at decision point j if $U_{ij}^A(\mathbf{X}_{it} \mid \mathbf{D}_{it}) \leq 0$ for $j \geq t$, in which case the final stage of the decision process, and the end of program participation, is defined as $t_{if} = j+1$. The final pattern of treatments taken, and thus the subset of program elements actually experienced by the DM, is revealed by the state of $\mathbf{D}_{it_{if}}$ at the stage, $t_{if} = j+1$, following the final decision point, j, at which the decision was made to conclude the course of treatments. A summary of the key elements of the decision process, and various possible outcomes of it, is presented in Table 3.1.

Table 3.1. Sequential Program Adherence Decision Process Summary

<u>Utility States</u>	<u>DM Decisions</u>
$U_{ij}^A(\mathbf{X}_{i1} \mid \mathbf{D}_{i1}) \leq 0, \ \forall j$	Does not participate in the treatment program.
$U_{ij}^A(\mathbf{X}_{i1} \mid \mathbf{D}_{i1}) > 0, \exists\, j$	Accepts some pattern of treatments, where the treatments ultimately taken are those for which $U_{ij}^A(\mathbf{X}_{ij} \mid \mathbf{D}_{ij}) > 0, \ j = 1,..., J$
$U_{ij}^A(\mathbf{X}_{ij} \mid \mathbf{D}_{ij}) > 0, \ \forall j$	Accept the complete set of prescribed treatments.
$U_{ij}^A(\mathbf{X}_{it} \mid \mathbf{D}_{it}) \leq 0, \ \forall j \geq t$	Stop treatments at decision stage j, so that $t_{if} = j+1$.

In the next section, we present an empirical implementation of the preceding general structure for the sequential adherence decision process that is consistent with all of the process features. The empirical implementation is computationally tractable and provides a relatively straightforward basis for interpreting and testing various hypotheses about the decision making processes of DMs, and the resulting patterns of treatments chosen and program experienced.

4.0 AN EMPIRICAL MODEL OF SEQUENTIAL ADHERANCE DECISIONS

In order to transition from the general decision process structure of the preceding section to a readily implementable empirical model, we utilize the commonly applied linear utility index in the model presentation, but the approach can be readily adapted to nonlinear utility index specifications, essentially by replacing the linear with a nonlinear systematic component wherever the index appears ahead, *mutatis mutandis*. The evolving stochastic component of the model accounts for the initial heterogeneity across DMs' perceptions of treatment experiences as well as the way anticipated utilities are sequentially updated.

Our empirical implementation is consistent with the basic random utility model (RUM) of Manski (1977), but represents a dynamic extension by accounting for DMs' learning about the value of treatments through their actual experience with treatments. In general concept, our approach bears some resemblance to the important recent work by Manski (2004, 2005) on random utility models subject to ambiguity. However, in that work decision makers must choose their actions at a specified time and cannot revise their choices once made. In contrast to our model ahead, Manski's DMs cannot undertake learning-by-doing and cannot accumulate empirical evidence to inform decisions, implying that each decision maker faces a single choice problem with predetermined information. The dynamics in that model emerge solely through social learning across successive cohorts and, unlike our approach, individual DMs do not themselves face dynamic choice problems.

4.1 Sequential Utility Specification: Rank One Heterogeneity

In this section we present a RUM specification that is based on a rank one stochastic representation of the initial heterogeneity of DMs' anticipated utilities. This is the more parsimonious stochastic specification of the model, but is still consistent with the features of the general sequential selection decision process presented in section 3. In section 4.2 ahead we indicate ways of generalizing the model, which also provides additional conceptual context for interpreting the implications of the more parsimonious specification.

At the first decision stage, $j=1$, the state of DM_i's $J\times 1$ anticipated utility vector for the *program* of treatments, having had no prior experience with the program, is assumed given by the following rank one distributed error formulation

$$\mathbf{U}_i^A\left(\mathbf{X}_{i1},\mathbf{D}_{i1}\right)=\mathbf{B}'\mathbf{X}_{i1}+\mathbf{c}_1\varepsilon_i\,. \tag{4.1}$$

In this specification, $\mathbf{B}=\left[\mathbf{B}_1\,|\,\mathbf{B}_2\,|\ldots|\,\mathbf{B}_J\right]$ is a $K\times J$ matrix[7], with $\mathbf{B}_\ell$ being the $K\times 1$ column vector of coefficients that determines the expected value of anticipated utility associated with the ℓ^{th} treatment for a DM characterized by $\mathbf{X}_{i1}$, as $\mathbf{B}_\ell{}'\mathbf{X}_{i1}$. The $K\times 1$ vector $\mathbf{X}_{i1}$ represents

[7] For simplicity of exposition, we are assuming that that the number of characteristics represented by $\mathbf{X}$ is the same across treatments. The model can be extended so that the number changes with *j* either by appropriately changing the dimension of the $\mathbf{B}_\ell$'s and corresponding $\mathbf{X}_i$'s, or else specifying the dimension of $\mathbf{X}_i$'s maximally and including zero values as appropriate vis-à-vis *t*.

$\mathrm{DM}_i\,'s$ relevant sociodemographic and/or treatment characteristics, and $\mathbf{c}_1$ is a $J \times 1$ vector of coefficients that interact with the random scalar $\varepsilon_i \sim F(\varepsilon)$ to determine the distribution of deviations, $\mathbf{c}_1\varepsilon_i$, across the J expected anticipated utility levels $\mathbf{B}'\mathbf{X}_{i1}$. We adopt a standard normal distribution for the deviation in what follows, i.e., $\varepsilon_i \sim N(0,1)$, but the model is general and can accommodate a myriad of alternative error distributions. As will become apparent, because of scale invariance, one can normalize the anticipated utility for the first treatment by setting $c_{11} = 1$ in the vector $\mathbf{c}_1$**,** with an appropriate scaled interpretation of $\mathbf{B}_1$.

The scalar random variable ε_i in (4.1) represents unobserved (by the econometrician) heterogeneity among the DMs, consistent with a RUM specification. We underscore that an implicit assumption in the current specification is that a rank one residual vector represents the initial heterogeneity across DMs' $J \times 1$ anticipated utility vectors adequately, i.e., $Cov\left(\mathbf{U}_i^A\right) = \mathbf{c}_1\mathbf{c}_1'$. This would be the case, for example, if unobserved factors affect the J anticipated utilities in some proportionate pattern. This specification would be applicable to a decision context such as that described by Manski and Tamer (2002), where the systematic component is specified well except for one important dimension of explanation that is not precisely measured – and in our application, not measured at all. Higher rank specifications which relax this implicit assumption and accommodate more complex patterns of heterogeneity are discussed in the next section.

In this distributed error formulation, the scalar random disturbance ε_i is distributed across the J utilities via the pattern indicated by the vector $\mathbf{c}_1$, which in unrestricted form utilizes $J-1$ freely varying parameters, given that the first coefficient is normalized to unity . A lesser number of free parameters would be required if the distribution of the error were specified according to a given parametric functional form, say $f\left(J;\boldsymbol{\theta}\right)\varepsilon_i$, for an appropriate $J\times 1$ vector function $f\left(J;\boldsymbol{\theta}\right)$. For example, a standard type of utility discounting framework represented by $f\left(J;a,r\right)=\left[ae^{-rt},\ t=1,\ldots,J\right]$ would involve only two parameters. More elaborate specifications could have the elements of $\mathbf{c}_1$ expressed as parametric functions of other explanatory variables.

The anticipated utility outcome $\mathbf{U}_i^A\left(\mathbf{X}_{i1},\mathbf{D}_{i1}\right)=\mathbf{B}'\mathbf{X}_{i1}+\mathbf{c}_1\varepsilon_i\leq\left[\mathbf{0}\right]$ coincides with a decision by DM_i not to participate in the treatment program. If DM_i decides to participate, she decides whether to take the first treatment via the decision rule (with the $c_{11}=1$ normalization imposed)

$$if\ \ U_{i1}^A\left(\mathbf{X}_{i1}\,|\,\mathbf{D}_{i1}\right)=\mathbf{X}_{i1}'\mathbf{B}_1+\varepsilon_i\begin{Bmatrix}>\\ \leq\end{Bmatrix}0\ then\begin{Bmatrix}accept\\ reject\end{Bmatrix}treatment\ 1 \tag{4.2}$$

and it follows that $D_{i2}\left[1\right]=I\left(U_{i1}^A\left(\mathbf{X}_{i1}\,|\,\mathbf{D}_{i1}\right)>0\right)$, where $I\left(A\right)$ is the standard indicator function that takes the value 1 when A is true.

The way in which the state vector of anticipated utilities is updated at decision point two and beyond depends on the outcomes of decisions made relative to accepting treatments. If a treatment is rejected, no updating occurs from that treatment since nothing new is experienced or learned. If a treatment is accepted, anticipated utilities are updated as a function of the treatment experience. Regarding the experientially derived updates, we assume that they are akin to those discussed by Block and Marschak (1960), whereby there are some aspects of taking treatments that cannot be anticipated, so that the DM lacked sufficient information to incorporate these aspects into her decision process.[8] Once experienced, this type of effect acts like a structural shift in the DM's treatment perceptions that can persist into the future, affecting future decisions. Anticipated utilities for remaining treatments are updated by amounts related to the treatment experience and knowledge gained based on the stage 1 decision[9], where at decision point t = 2,

$$\begin{bmatrix} \mathrm{U}_{i2}^{A}\left(\mathbf{X}_{i2},\mathbf{D}_{i2}\right) \\ \mathrm{U}_{i3}^{A}\left(\mathbf{X}_{i2},\mathbf{D}_{i2}\right) \\ \vdots \\ \mathrm{U}_{iJ}^{A}\left(\mathbf{X}_{i2},\mathbf{D}_{i2}\right) \end{bmatrix} = \begin{bmatrix} \mathbf{X}_{i2}'\mathbf{B}_{2} \\ \mathbf{X}_{i2}'\mathbf{B}_{3} \\ \vdots \\ \mathbf{X}_{i2}'\mathbf{B}_{J} \end{bmatrix} + \mathbf{D}_{i2}\left[1\right]\begin{bmatrix} c_{22} \\ c_{32} \\ \vdots \\ c_{J2} \end{bmatrix} v_{i1} + \begin{bmatrix} c_{21} \\ c_{31} \\ \vdots \\ c_{J1} \end{bmatrix}\varepsilon_{i} \qquad (4.4)$$

The random scalar $v_{i1} \sim N\left(0,1\right)$ in (4.4) interacts with the vector $\left[c_{22},c_{32},\ldots,c_{J2}\right]'$ to represent a distribution of updates to anticipated utilities for treatments $j \geq 2$ associated with experiencing treatment 1. In general, the updating of remaining treatments is given by

$$\begin{bmatrix} \mathrm{U}_{it}^{A}\left(\mathbf{X}_{it},\mathbf{D}_{it}\right) \\ \mathrm{U}_{i,t+1}^{A}\left(\mathbf{X}_{it},\mathbf{D}_{it}\right) \\ \vdots \\ \mathrm{U}_{iJ}^{A}\left(\mathbf{X}_{it},\mathbf{D}_{it}\right) \end{bmatrix} = \begin{bmatrix} \mathbf{X}_{it}'\mathbf{B}_{t} \\ \mathbf{X}_{it}'\mathbf{B}_{t+1} \\ \vdots \\ \mathbf{X}_{it}'\mathbf{B}_{J} \end{bmatrix} + \sum_{\ell=2}^{t}\mathbf{D}_{i\ell}\left[\ell-1\right]\begin{bmatrix} c_{t\ell} \\ c_{t+1,\ell} \\ \vdots \\ c_{J\ell} \end{bmatrix} v_{i,\ell-1} + \begin{bmatrix} c_{t1} \\ c_{t+1,1} \\ \vdots \\ c_{J1} \end{bmatrix}\varepsilon_{i} \;\; for\ t \geq 2 \qquad (4.5)$$

where the random scalar $v_{i,\ell-1} \sim N\left(0,1\right)$ interacts with the vectors $\left[c_{t\ell},c_{t+1,\ell},\ldots,c_{J\ell}\right]'$ for $\ell = 2,\cdots,t,$ to define a distribution of updates to anticipated utilities for treatments $j \geq t$

---

[8] This assumption is consistent with rational behavior. For example, two patients assigned the same drug treatment might realize different utilities simply because one likes cherries and the other does not, yet neither knew the drug was cherry flavored. In an example more pertinent to our application later in this paper, a patient may relate better to a counseling session provided by a female presenter than a male, and unknown to the patient, the program leader turns out to be female.

[9] The distributed error characterization of what is learned, besides the possible update from $\mathbf{X}_{i,\ell-1}$ to $\mathbf{X}_{i\ell}$, by attending a session does not carry with it the implicit assumption that underlies the rank 1 distribution of the initial disturbance term, ε_i. Instead, v_{i1} represents a single unanticipated experiential shock, again unobserved by the econometrician, that can have different impacts on future anticipated utilities. The elements in the vector $\mathbf{c}_{.\ell}$ represent these different possible impacts.

associated with experiencing some combination of treatments $j = 1, \ldots t-1$. We assume that

$v_i = \left[v_{i1}, \ldots, v_{i,J-1} \right]' \sim N\left(\mathbf{0}, \mathbf{I}_{J-1}\right)$, although again the assumption of normality is not necessary.

The entire evolution of the anticipated utilities that determine DM_i's treatment decisions can be represented in a recursive matrix format as follows:

$$\begin{bmatrix} \mathrm{U}_{i1}^A\left(\mathbf{X}_{i1}, \mathbf{D}_{i1}\right) \\ \mathrm{U}_{i2}^A\left(\mathbf{X}_{i2}, \mathbf{D}_{i2}\right) \\ \mathrm{U}_{i3}^A\left(\mathbf{X}_{i3}, \mathbf{D}_{i3}\right) \\ \vdots \\ \mathrm{U}_{iJ}^A\left(\mathbf{X}_{iJ}, \mathbf{D}_{iJ}\right) \end{bmatrix} = \begin{bmatrix} \mathbf{X}_{i1}'\mathbf{B}_1 \\ \mathbf{X}_{i2}'\mathbf{B}_2 \\ \mathbf{X}_{i3}'\mathbf{B}_3 \\ \vdots \\ \mathbf{X}_{iJ}'\mathbf{B}_J \end{bmatrix} + \left(\begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ c_{21} & c_{22} & 0 & \cdots & 0 \\ c_{31} & c_{32} & c_{33} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ c_{J1} & c_{J2} & c_{J3} & \cdots & c_{J,J} \end{bmatrix} \odot \begin{bmatrix} 1 & I_{i1} & I_{i2} & \cdots & I_{i,J-1} \end{bmatrix} \right) \begin{bmatrix} \varepsilon_i \\ v_{i1} \\ v_{i2} \\ \vdots \\ v_{i,J-1} \end{bmatrix} \quad (4.6)$$

where $I_{ij} = \begin{Bmatrix} 1 \\ 0 \end{Bmatrix}$ *if* $\begin{Bmatrix} \mathrm{U}_{ij}^A\left(\mathbf{X}_{ij}, \mathbf{D}_{ij}\right) > 0 \\ otherwise \end{Bmatrix}$, and $\begin{bmatrix} \varepsilon_i \\ v_{i1} \\ v_{i2} \\ \vdots \\ v_{i,J-1} \end{bmatrix} \sim N\left(\mathbf{0}, \mathbf{I}_J\right)$. We will have use ahead for the compact representation of the vector equation (4.6) given by

$$\mathrm{U}_i^A\left(\mathbf{X}_i, \mathbf{D}_i\right) = \mathbf{X}_i' vec\left(\mathbf{B}\right) + \left(\mathbf{C} \odot \mathbf{\Phi}_i'\right) \mathbf{V}_i \quad (4.7)$$

where $\mathbf{X}_i \equiv \begin{bmatrix} \mathbf{X}_{i1} & & & \\ & \mathbf{X}_{i2} & & \\ & & \ddots & \\ & & & \mathbf{X}_{iJ} \end{bmatrix}$, and the matrices in last additive term in (4.7) are defined in a way that is apparent upon making one-by-one associations with the appropriate vectors or matrices in (4.6), and $\odot$ is the generalized Hadamard (elementwise) product operator. If at decision point t, $\mathbf{U}_{ij}^A\left(\mathbf{X}_{it}, \mathbf{D}_{it}\right) \leq 0 \ \forall j \geq t$, then the treatment program is terminated, updating stops, and this event would be reflected in an appropriate set of ones and zeros for the indicators in the vector $\mathbf{\Phi}_i' = \begin{bmatrix} 1 & I_{i1} & I_{i2} & \cdots & I_{i,J-1} \end{bmatrix}$.

4.2 Sequential Utility Specifications: Higher Rank Heterogeneity and Other Generalizations

At the expense of added parametric, functional and computational complexity, the model can be generalized in a number of potentially useful ways. These include

1) higher rank stochastic error structures for incorporating more diverse and complicated patterns of heterogeneity underlying the initial period RUM specification,

2) allowing the coefficient vectors defining the distributed error structure to vary as a function of other observable factors, and

3) allowing for nonlinear relationships with factors affecting the level of the anticipated utility levels.

All but the higher rank stochastic structure can be accommodated in relatively straightforward and expected ways that require little additional discussion, although one may encounter additional computational and identification issues that would need to be addressed when estimating the model. A condensed depiction of the incorporation of 2) and 3) is given by the following generalization of the parsimonious model (4.7):

$$\mathrm{U}_i^A\left(\mathbf{X}_i, \mathbf{D}_i\right) = \boldsymbol{\mathcal{U}}\left(\mathbf{X}_i; \mathbf{B}\right) + \left(\mathbf{C}\left(\mathbf{X}_i; \boldsymbol{\theta}\right) \odot \mathbf{\Phi}_i'\right) \mathbf{V}_i \tag{4.8}$$

where now $\boldsymbol{\mathcal{U}}(\bullet)$ and $\mathbf{C}(\bullet)$ are, respectively, general vector and matrix functions of sociodemographic and/or treatment characteristics.

In order to motivate higher rank error structures for depicting RUM heterogeneity, first consider a general rank ℓ representation of a rank $J > \ell$ covariance structure $\mathbf{\Sigma}$. It is well known that one can generate an approximation to a higher rank positive semidefinite matrix $\mathbf{\Sigma}$, in the sense of minimizing the Frobenius norm, by using the appropriate first ℓ components of a singular value decomposition corresponding to the largest singular values, s_i, as $\Sigma \approx \sum_{j=1}^{\ell} s_j \mathbf{v}_j \mathbf{v}'_j$, where $\mathbf{v}_j$ is the left singular vector, and $\mathbf{v}'_j$ is then the right singular vector by virtue of the symmetry and positive semidefiniteness of $\mathbf{\Sigma}$. The best possible rank one approximation to $\mathbf{\Sigma}$, which minimizes the Frobenius norm amongst all possible rank one matrices, is given by $s_1 \mathbf{v}_1 \mathbf{v}'_1$ with s_1 being the largest singular value. The approximation is progressively improved, but with norm improvements progressively diminishing, when ℓ is increased in the representation $\Sigma \approx \sum_{j=1}^{\ell} s_j \mathbf{v}_j \mathbf{v}'_j$. Thus motivation for the parsimonious rank one error structure in section 4.1 follows from viewing $\mathbf{c}_1$ as a potential parameterization of the best rank one approximation to $\mathbf{\Sigma}$, i.e., $\mathbf{c}_1 = \sqrt{s_1}\mathbf{v}_1$. However, note that as ℓ increases, the number of entries in the singular vectors, if viewed as parameters, eventually exceed the maximal number of unknown parameters in $\mathbf{\Sigma}$, i.e. for large enough ℓ, it follows that $\ell J > J(J+1)/2$.

Recent work by Harbrecht, Peters, and Schneider (2010) motivate a new low rank approximation methodology for symmetric positive semidefinite matrices that is consistent with a progressively more dense parameterization, and a correspondingly higher rank and increasingly more accurate stochastic approximation that ends with a perfect representation of the matrix with no more than $J(J+1)/2$ unique parameters employed, depending on the rank of $\mathbf{\Sigma}$. In particular, the use of the first ℓ factors of a Cholesky decomposition provides a progressively more accurate representations of a symmetric positive semidefinite matrix, as

$\Sigma \approx \sum_{j=1}^{\ell} \xi_j \xi_j'$, where the ξ_j's are the successive column vectors in the triangular Cholesky matrix factorization of $\mathbf{\Sigma}$. Using the foregoing as motivation, the initial RUM error structure, $\mathbf{c}_1 \varepsilon_i$, shown in (4.5) for the rank one model, could be generalized to $\mathbf{C}_1 \boldsymbol{\varepsilon}_i = \sum_{j=1}^{\ell} \mathbf{C}_{1j} \boldsymbol{\varepsilon}_{ij}$, where now $\mathbf{C}_1$ is a $J \times \ell$ coefficient matrix with zeros in every position (i, j) for which $j > i$, $\mathbf{C}_{1j}$ is the j^{th} column of $\mathbf{C}_1$, and $\boldsymbol{\varepsilon}_i \sim N(\mathbf{0}, \mathbf{I}_\ell)$ (or other choice of $\ell-$variate distribution).

The choice of appropriate rank ℓ, as well as which entries in the $J \times \ell$ coefficient matrix $\mathbf{C}_1$ are pertinent for representing heterogeneity, can be investigated via likelihood ratios or other standard testing methodology. In this way, additional flexibility through higher rank stochastic heterogeneity can be incorporated into model (4.6) as appropriate.

4.3 Maximum Likelihood Estimation

We concentrate here on the parsimonious selection decision model presented in section 4.1, which has some unique stochastic features that facilitate devising an estimation approach for the unknown parameters of the model. In specifying the likelihood function for ML estimation, the rank one stochastic structure introduces integration dimension-reducing characteristics for any DM observations that are associated with less than complete adherence to the full complement of J treatments in the program. In particular, the probability of any combination of treatment acceptances and rejections, which is characterized by a collection of lower bounded (>0) and upper bounded (≤ 0) events applied to the anticipated utilities in (4.6), involve at most a J-dimensional multivariate normal integral, and if any m of the first

$J-1$ treatments are missed, the dimension of the multivariate normal distribution (i.e., the dimension of the right most stochastic vector in (4.6)) is reduced correspondingly to $J-m$.

Given the preceding dimension reduction, one can proceed to assign the probability of any combination of sequential treatment acceptances and rejections in at least two ways. One is to transform the J anticipated utility inequalities in terms of the corresponding $J-m$ dimensional random vector on the RHS of (4.6), defined by the nonzero entries in the vector[10] $\mathbf{V}_i \odot \mathbf{\Phi}_i$, and then calculate the probability of the event defined by the inequalities, resulting in the integration of a $J-m$ dimensional normal distribution. An alternative approach, which will result in precisely the same event probability, is to use the J-dimensional, rank $J-m$ linear transformation of the disturbance vector in (4.6), $\mathbf{Z} = \left(\mathbf{C} \odot \mathbf{\Phi}_i'\right)\mathbf{V}_i$, which defines a $J \times 1$ *singular* normal random vector $\mathbf{Z}$ of rank $J-m$, and calculate the probability of the event defined by the J inequalities imposed directly on $\mathbf{Z}$. It will be seen below, in implementing ML estimation for the decision model in 4.1, that using the singular normal approach induces some substantial computational advantages in the calculation of the event probabilities.

Assume that there are sample observations available on *n* DMs who have made decisions regarding accepting treatments in a program consisting of a sequence of J treatments. The pattern of acceptances and rejections of treatments will determine which $\mathrm{U}_{ij}^{A}\left(\mathbf{X}_{ij}, \mathbf{D}_{ij}\right)s$ are >0 (acceptances) and which are ≤ 0 (rejections). Through an appropriate reflection (around the origin) of random variables and associated events relating to the DM's response, all DM's decision can be cast in terms of a (possibly singular) multivariate normal CDF

[10] Note that $\left(\mathbf{C} \odot \mathbf{\Phi}_i'\right)\mathbf{V}_i = \mathbf{C}\left(\mathbf{V}_i \odot \mathbf{\Phi}_i\right)$

value. To accomplish this, first define the following "reflection indicator", where recall that $\mathbf{D}_{it_f}$ is a vector of 1's and 0's indicating the final decisions made by DM_i regarding all treatment opportunities:

$$\mathbf{S}_i = \mathbf{1}_J - 2\mathbf{D}_{it_f} \tag{4.9}$$

The j^{th} entry in the vector $\mathbf{S}_i$ has a value of -1 if DM_i chose to adhere to the j^{th} treatment and has a value of 1 otherwise.

It follows that the DM's treatment choice event, and its probability of occurrence, is given by the following CDF representation (recall (4.6) and (4.7)):

$$\begin{aligned} F_i\left(-\left(\mathbf{X}_i' vec\left(\mathbf{B}\right)\right)\odot \mathbf{S}_i \mid \boldsymbol{\Sigma}_i\left(\mathbf{C}\right)\right) &= P\left(\left(\left(\mathbf{S}_i \odot \mathbf{C}\right)\odot \boldsymbol{\Phi}_i'\right)\mathbf{V}_i \le -\left(\mathbf{X}_i' vec\left(\mathbf{B}\right)\right)\odot \mathbf{S}_i\right) \\ &= \int\cdots\int_{\mathbf{W}\le -\left(\mathbf{X}_i' vec\left(\mathbf{B}\right)\right)\odot \mathbf{S}_i} N\left(\mathbf{W};\, \mathbf{0}, \boldsymbol{\Sigma}_i\left(\mathbf{C}\right)\right)\mathbf{dW} \end{aligned} \tag{4.10}$$

which is the value of a possibly singular multivariate normal CDF, with mean vector **0** and covariance matrix $\boldsymbol{\Sigma}_i\left(\mathbf{C}\right) = \left(\left(\mathbf{S}_i \odot \mathbf{C}\right)\odot \boldsymbol{\Phi}_i'\right)\left(\left(\mathbf{S}_i \odot \mathbf{C}\right)\odot \boldsymbol{\Phi}_i'\right)'$. The log-likelihood function implied by (4.10) is then defined by

$$\ell\left(\boldsymbol{\beta}, \mathbf{C} \mid \mathbf{D}_{t_f}, \mathbf{X}\right) = \sum_{i=1}^{n} \ln\left(L_i\left(\boldsymbol{\beta}, \mathbf{C} \mid \mathbf{D}_{it_f}, \mathbf{X}_i\right)\right) = \sum_{i=1}^{n} \ln\left(F_i\left(-\left(\mathbf{X}_i' vec\left(\mathbf{B}\right)\right)\odot \mathbf{S}_i \mid \boldsymbol{\Sigma}_i\left(\mathbf{C}\right)\right)\right) \tag{4.11}$$

and the maximum likelihood estimate of the parameters of the sequential decision model is given by

$$\left\{\hat{\boldsymbol{\beta}}_{ML}, \hat{\mathbf{C}}_{ML}\right\} = \underset{\mathbf{B},\mathbf{C}}{\arg\max}\left\{\sum_{i=1}^{n} \ln\left(F_i\left(-\left(\mathbf{X}_i' vec\left(\mathbf{B}\right)\right)\odot \mathbf{S}_i \mid \boldsymbol{\Sigma}_i\left(\mathbf{C}\right)\right)\right)\right\} \tag{4.12}$$

4.4 Computational Issues

Common to all models that require integration of a multivariate normal distribution, the dimensionality of the integration problem becomes an issue. As the dimension increases beyond 4, numerical integration of the multivariate normal integral becomes progressively much slower, more inaccurate, and eventually intractable. One then must resort to some form of stochastic Monte Carlo-type integration approach, or else use some recursive integration method based on iterative conditioning, often with approximations introduced as well (for example, see Genz (1992), Hayter (2006), and Richard and Zhang (2007)). One of the attractive features of the current distributed error model is that the dimension of the integration required decreases commensurate with the number of treatments missed, and so integration in the full J dimensions is generally not required for all *n* sample observations, such as in the empirical application we examine ahead in section 5. However, relatively high dimensional integration remains for many sample observations, and there is the added issue of the singularity of the normal distributions for some of the sample observations.

We have examined a number of methods for calculating the *n* probabilities required for composing the likelihood function in (4.11)-(4.12). These included the use of recursive importance sampling, kernel density approximations, Genz's well-known numerical approaches for calculating high dimensional normal integrals specialized to singular normal distributions (see Genz and Kwon (2000)), and direct simulation of the multivariate normal integral for both the singular and non-singular covariance structure cases. The latter direct approach was substantially more efficient for calculating the probability elements of the likelihood function, and for calculating the ML estimates themselves. The approach was facilitated by using four threads of parallel processing and liberal use of generalized Hadamard operators (to avoid the need for time-consuming looping) on a PC with a quad-core CPU, running at 3.0 GHz, having 12 GB of memory, and the computer program was written in GAUSS™ version 11. The random number generator used was GAUSS's internal **rndKMn** KISS+Monster program designed by George Marsaglia[11], combined with Kinderman and Ramage's (1976) acceptance-rejection algorithm[12]. Additional details regarding the direct simulation approach, and it computational advantages, are provided in the Appendix.

4.5 A Sampling Experiment

In order to illustrate the application of the estimation methodology within a context that provides an illustration of the repeated sampling characteristics of the approach, as well as observations on the potential time-to-convergence characteristics of the method, we provide a MC simulation example here for a hypothetical program consisting of $J = 4$ treatments. The parameter values underlying the decision model (4.6) were set as follows:

$$\mathbf{B} = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 2 & 3 & 4 & 5 \\ 3 & 4 & 5 & 6 \end{bmatrix} \text{ and } \mathbf{C} = \begin{bmatrix} 1 & & & \\ .75 & .5 & & \\ .5 & .25 & .25 & \\ .25 & .125 & .125 & .0625 \end{bmatrix} \qquad (4.15)$$

and sample decision outcomes from (4.6) were generated using the decision rules and distributional specifications defined previously[13]. The $\mathbf{X}_i$'s were generated using *iid* outcomes

---

[11] For more information on the generator, see http://www.Aptech.com/random.

[12] The software underlying the results reported in this paper is available from the authors upon request.

[13] All simulated random variable outcomes were again generated using the KISS+Monster algorithm designed by Marsaglia, as noted in section 4.4.

from a $N\left(\mathbf{0}_3, \mathbf{I}_3\right)$ distribution and were held constant across treatments to mimic characteristics of the empirical application presented ahead in section 5. A total of 1000 repetitions for the probability calculation and the ML estimation methodology were used in the analysis of repeated sampling behavior for simulated sample sizes of $n = 50, 100, 250, 500$, as well as to gauge time-to-convergence characteristics of the approach. The maximization of the likelihood function (4.12) was accomplished by applying the Nelder-Mead Direct Search Optimization approach, as outlined by Jacoby, Kowalik, and Pizzo (1972).

A visual depiction of the overall results of the experiment in terms of the estimated average mean square error[14] of the parameters, and the average convergence time, is given in Figure 1.[15] It is evident from the figure that convergence time for obtaining the ML estimates increases at an increasing rate as the size of the data sample increases. However, even for a size 500 sample, the average convergence time for each of the 1000 estimation problems was notably less than one minute. The average MSE of the parameter estimates decreases monotonically with increasing sample size, as one would expect from a consistent ML estimator. Note that the squared bias component of the MSE was relatively small compared to the variance component, and so the decline in MSE observed in the figure is closely related to decreasing variation in the estimates of the model parameters. Overall, the method proved to

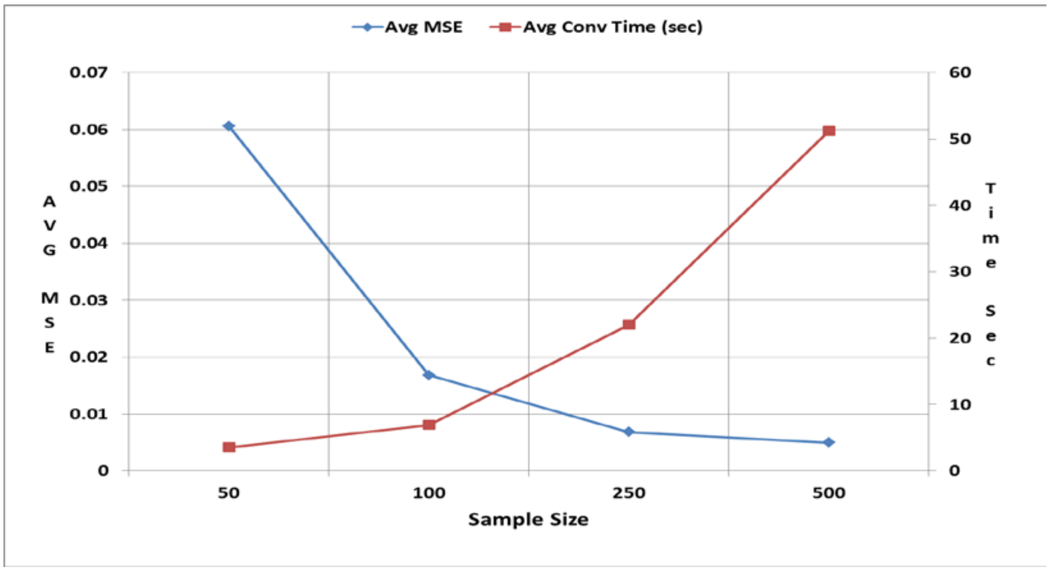
---

[14] The average MSE was calculated as $\dfrac{\sum_{\ell=1}^{M} \begin{pmatrix} \hat{\mathbf{B}}_{ML} - \mathbf{B} \\ \hat{\mathbf{C}}_{ML} - \mathbf{C} \end{pmatrix}' \begin{pmatrix} \hat{\mathbf{B}}_{ML} - \mathbf{B} \\ \hat{\mathbf{C}}_{ML} - \mathbf{C} \end{pmatrix}}{MK}$ where K is the total number of parameters that are being estimated.

[15] Detailed results relating to the sampling performance for each parameter is available from the authors upon request. In general, there were no remarkable differences in performance between estimates of $\mathbf{B}$ or $\mathbf{C}$, or between the estimates within each of the two parameter matrices.

be a rather accurate and reasonably fast estimation approach for the illustrative application examined here.

Figure 1. Average MSE and Convergence Time



5.0 Participation in State-Run Strengthening Families Programs

In order to provide a substantive empirical example of the methodology, the preceding model was applied to data pertaining to the Strengthening Families Program for Parents and Youth 10-14 in the states of Oregon and Washington. The Strengthening Families Program (SFP) is a nationally and internationally recognized parenting and family strengthening program that is designed to be delivered in local communities for groups of participants from 7-12 families. Family participants are asked to attend SFP once a week for seven weeks and participate in educational activities that bring parents and their children together in learning environments designed to strengthen families through improved family communication, parenting practices, and parents' family management skills.

5.1 Data

The data consisted of 860 parents who enrolled for one of the 94 SFP cycles in Oregon and Washington State in 2008 and 2009. The majority of program participants (69%) were identified as female, while 26% were identified as male. The remaining 5% did not report their gender. Thirty-three percent of the participants identified themselves as Hispanic/Latino and 57% were identified as White. The number of participants identifying themselves as belonging to one of the three minority groups, African-American, American Indian/Alaska Native, and other or multiple race/ethnicity was approximately 2% for each category, while 5% did not report their race/ethnicity. The average age of a participant in the sample was 39.3 years with a standard deviation of 8.4 years. The median age was 35 years. Over 72% of the sample observations were from Oregon, with the remainder from Washington. Only one-third of the participants attended all 7 sessions, and 74% attended at least 5 of the 7 sessions.

5.2 Results Overview

The empirical model described in section 4.1, with $J = 7$, was applied using participant demographics representing gender, ethnicity and age. In particular, the covariates included indicator variables which equaled 1 if gender was male, gender was not reported, African American, Hispanic/Latino, Native American, other or multiple ethnicity, or race/ethnicity unreported. The actual age of the participant was also included. Descriptive statistics for the demographic variables used as covariates in the model, as well as for session attendance, are presented in Table I. Note that because only sociodemographic information pertaining to participants was available, participant characteristics remained constant across treatments in this application.

The results of ML estimation[16] are reported in Table 2 for the $\mathbf{B}$ matrix and in Table 3 for the **C** matrix appearing in (4.6). The results in Table 2 determine the expected anticipated utility components, $\mathbf{X}_i' vec(\mathbf{B})$, of the decision process, which remain constant over a participant's J decision points because participant characteristics are constant over the decision horizon, as noted above. This implies that in the current application, it is appropriate to interpret $\mathbf{X}_i' vec(\mathbf{B})$ as representing the mean anticipated utilities for the J treatments prevailing at the *first* decision point, as well as the central tendencies for anticipated utilities at future decision points. The results in Table 3 define, via $(\mathbf{C} \odot \mathbf{\Phi}_i')\mathbf{V}_i$, the dynamic stochastic process by which anticipated utilities of the participants are updated via the choice of adherence to, and experience of various sessions in the program. The large majority of the estimated parameters in each of the matrices were statistically significant at conventional .01 or .05 levels of type I error[17]. A likelihood ratio test of the joint significance of all explanatory factors in the model beyond the intercepts was equal to 768, which far exceeded the .01 level critical value of Chi-square(0.01, 83)=115.

The mean anticipated utilities from the treatment sessions vary a great deal across ethnic identities, gender and age. The results raise questions for program providers regarding how and why the program content provided in the various sessions had such a heterogeneous

[16] The ML estimates required 5 minutes and 1 second to converge on the quad core PC described in the MC section.
[17] Standard deviations of the estimated parameters were calculated via bootstrap resampling of the observed sample observations.

influence on the propensity of the various types of participants to attend specific sessions. For example, based on the estimated anticipated utilities, all participants who belonged to a minority ethnic group were less likely to attend the first session than whites. Blacks were less likely than whites to attend each of the sessions, except the last. Hispanics were less likely to attend the first and the penultimate session than whites, but were more likely to attend each of the other sessions. Older participants were more likely to attend the third, fourth and the sixth session than younger participants. Male participants were less likely to attend the third and fifth sessions relative to female participants, but were more likely to attend the fourth session. Overall, the results for the $\mathbf{B}$ coefficients suggest that participants choose diverse patterns of session adherence and hold varying intentions of attending late sessions in the treatment sequence, depending on age, ethnicity and gender.

Regarding estimates of the **C** matrix entries, one model specification check is whether coefficients on the residuals associated with sessions attended (i.e., the $\nu_{i\ell}$'s) are all positive. Negative coefficients imply that a positive experiential adjustment to utility derived from session attendance negatively affects anticipated utilities of future sessions, and vice versa, which would be counterintuitive for the program under study. Table 3 reveals that all of the elements of the **C** matrix were estimated to be positive.

In general, the estimated **C** matrix suggests that session experiences have significant impacts on the perceived value of subsequent program sessions, and initial anticipated utilities of sessions can be significantly altered by session experiences. In particular, based on anticipated utilities implied by the estimated model, many participants who attend the program do not initially intend to adhere to all seven sessions. For example, an average female Hispanic participant enrolls in SFP with the intent of attending the first five program sessions, and an average female white or African American participant does not initially look forward to continuing the program beyond the fourth session. In both cases, it is the negative or positive experientially-derived increments to utility from earlier sessions attended that drives participants to change these plans.[18]

Note that all but two of the coefficient estimates presented in Table 3 are statistically significant at conventional levels of type 1 error. Interestingly, one of the insignificant coefficients relates to the effect that experiencing the first treatment has on attending the second session, suggesting that first session attendees may prefer to 'wait-and-see' rather than over-react to their first impressions regarding the treatment program. Moreover, the sixth session also has no statistically significant impact on the decision to attend the final session, suggesting that even a poor experience in the penultimate session of the program would not act as a deterrent to participating in the subsequent session that completes the program.

The complexity of attendance decisions is substantial beyond the third session and the decision of a given participant to attend a session depends not only on gender, ethnicity and age but also on which past sessions were attended. By accounting for experiential updating, the estimated model predicts the level of compliance at the first four sessions very closely, but underpredicts the level of compliance thereafter. The model also predicts correctly the gradually increasing program attrition until the sixth session as well as the elevated level of compliance at the last session.

[18] Predicted anticipated utilities were computed from the estimated parameters and age was evaluated at its mean (39.3 years). Hence, the predicted anticipated utility for females follows the equation $U_{ij}^{A} = cons_{ij} + \beta_{3j} Black_i + \beta_{4j} Hisp_i + \beta_{10j} \overline{Age}$ where the j subscript indicates the session in question and the i subscript identifies the individual and Age was evaluated at its mean value.

6.0 Summary

We presented a relatively straightforward, parsimonious, and empirically tractable approach, rooted in a consistent behavioral context, for modeling heterogeneous sequential treatment decisions across DMs. The basic model rationalizes observed patterns of treatment adherence using a model of choice that accounts for observable factors influencing choices as well as accumulated learning derived from experiencing past treatments. The model can be applied in a variety of decision environments involving complex combinations of adherence patterns, and allows DMs to repeatedly discontinue and rejoin the treatment process, a feature not possible, or at least not a natural outcome of previous models in the literature.

The empirical illustration of the methodology suggested important behavioral aspects of the decision making process of participants in family strengthening programs, and illustrated the substantive types of information available from the model. In particular, the model explains the substantial number of dropouts from the program, and also helps to identify the types of participants who are more likely to dropout. The estimation proceeded unconstrained, and the result was that there was substantial heterogeneity in terms of the probability of session adherence, both in terms of the effects of participant sociodemographics, as well as the effects of experiencing various configurations of session experiences. In our application, we assumed normality and a linear functional form for the utility index, and data was available only on participant characteristics, and not on session characteristics. As we noted earlier, alternative distributional and index specifications can be accommodated, and additional data on factors affecting the adherence decision environment can be taken into account within the general model framework. The estimated heterogeneity in adherence propensities relating to the SPF program suggests, in fact, that there are program session idiosyncrasies, as well as potential additional participant characteristics that may be affecting participants' decisions to attend program sessions.

Based on the presentation and application of the basic sequential decision model in this paper, additional empirical work is now being pursued in which data on treatment characteristics are available and DM characteristics are time-varying. The need for, as well as the identification and effects of incorporating higher rank stochastic heterogeneity is also a focus of future research. Overall, the approach provides a well-defined template for representing and empirically analyzing the complex decision process underlying a series of sequential decisions that are informed by sequential learning, and for which repeated entry and exit is allowed.

Appendix: Computation Details

The essential elements of the direct simulation approach for calculating the likelihood function are straightforward, and contribute to its programming and computational efficiency. In brief, first a $(J \times nM)$ matrix of *iid* standard normal random variable outcomes, $\mathbf{Z}$, is generated, where J is the number of treatments, n is the data sample size, and M is the number of Monte Carlo repetitions on which the multivariate normal probability calculation is based (e.g., $M = 10,000$ repetitions was used in this paper). Defining $\mathbf{\Phi} \equiv [\mathbf{\Phi}_i, i = 1, ..., n]$ and $\mathbf{S} \equiv [\mathbf{S}_i, i = 1, ..., n]$, the $\mathbf{Z}$ outcomes are then transformed as

$$\mathbf{W} = \left[\mathbf{C}\left(\mathbf{Z} \odot \left(\mathbf{1}'_M \otimes \mathbf{\Phi}\right)\right)\right] \odot \left(\mathbf{1}'_M \otimes \mathbf{S}\right) \tag{A.1}$$

so that the $(J \times nM)$ matrix $\mathbf{W}$ contains M successive $(J \times n)$ blocks of random variable outcomes, where the n columns within each of these blocks define the n outcomes of $\left((\mathbf{S}_i \odot \mathbf{C}) \odot \mathbf{\Phi}'_i\right)\mathbf{V}_i$, $i = 1, ..., n$ respectively. Then $\mathbf{W}$ is transformed into a matrix of binary outcomes by applying an appropriate column vector-wise indicator function operator $\mathbf{I}(\mathbf{A} \le_c \mathbf{H})$, which for any two $a \times h$ matrices, $\mathbf{A}$ and $\mathbf{H}$, returns a $h \times 1$ vector of 1's and 0's, with a value of 1 in the ℓ^{th} position indicating that *all* elements in the ℓ^{th} column of $\mathbf{A}$ were $\le$ the corresponding elements in the ℓ^{th} column of $\mathbf{H}$. In particular, letting $\mathcal{XB} \equiv \left[\mathbf{X}'_1 vec(\mathbf{B}) \mid \mathbf{X}'_2 vec(\mathbf{B}) \mid \cdots \mid \mathbf{X}'_J vec(\mathbf{B})\right]$, the transformation is defined by

$$\mathcal{I} = \mathbf{I}\left(\mathbf{W}. \le_c \left[\mathbf{1}'_M \otimes \left(-\mathcal{XB} \odot \mathbf{S}\right)\right]\right) \tag{A.2}$$

which yields an $(nM \times 1)$ binary vector containing M vertically concatenated $n \times 1$ vectors, the elements of which indicating when a simulated outcome of $((\mathbf{S}_i \odot \mathbf{C}) \odot \mathbf{\Phi}_i')\mathbf{V}_i$ fully satisfied (= 1) or did not fully satisfy (= 0) its respective upper bound $-(\mathbf{X}_i' vec(\mathbf{B})) \odot \mathbf{S}_i$, for $i = 1, \ldots, n$. Then the $n \times 1$ vector of estimated probabilities corresponding to the n sampled DM outcomes are straightforwardly obtained using the **reshape**(**A**,r,c)[19] operator, as

$$\mathbf{P}' = \mathbf{1}_M' \, \mathbf{reshape}(\mathcal{J}, M, n) / M \ . \tag{A.3}$$

The estimated probabilities in (A.3), which will depend on the values of the parameters $\mathbf{B}$ and $\mathbf{C}$, can be inserted into (4.11) to provide an estimated value of the likelihood function on which to base ML estimates of the model parameters.

In the actual software implementation of the above probability estimation methodology, the steps implied by (A.1) and (A.2) were separated into four simultaneous parallel subsample calculations, allocating $M/4$ repetitions to each, to exploit the capability of multiple core PCs[20], and the results were then reassembled to calculate the final step (A.3). Note that steps (A.1) – (A.3) literally require only 3 lines of programming code, suggesting the substantial simplicity and efficiency with which the calculations can be made, despite the relatively complicated heterogeneous nature of the $(\mathbf{C} \odot \mathbf{\Phi}_i')$ matrices and the numerous

[19] The **reshape**(**A**,r,c) operator creates a matrix from the vector or matrix **A** by using the elements of **A** to fill out an $r \times c$ matrix, with all elements stored in row-major order. For example, if $\mathbf{A} = [1 \quad 2 \quad 3 \quad 4 \quad 5 \quad 6]'$, then $\mathbf{reshape}(\mathbf{A}, 3, 2) = \begin{bmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \end{bmatrix}$.

[20] The number of parallel threads can, of course, be increased depending on the number of cores imbedded in the PC's CPU.

reflection operations needed to account for the varying patterns of program adherence decisions across DMs.

References

Angrist, J.D. and G.W. Imbens (1995), “Two-Stage Least Squares Estimation of Average Causal Effects in Models with Variable Treatment Intensity”, *Journal of the American Statistical Association,* 90(430), pp. 431-442.

Barrios, J.A. (2004), “Generalized sample selection bias correction under RUM”, *Economics Letters,* 85(1), pp. 129-132.

Basu, A., J.J. Heckman, S. Navarro-Lozano, S. Urzua (2007), “Use of instrumental variables in the presence of heterogeneity and self-selection: an application to treatments of breast cancer patients.” *Health Economics,* 16(11), pp. 1133–1157.

Block and Marschak (1960), “Random orderings and stochastic theories of response,” in Olkin (ed.) *Contributions to Probability and Statistics*, Stanford University Press,.

Chib, S. and L. Jacobi (2008), *“*Analysis of treatment response data from eligibility designs”, *Journal of Econometrics,* 144(2), pp. 465-478.

Crouchley, R. and M. Ganjali (2002), “The common structure of several models for non-ignorable dropout”, *Statistical Modelling*, 2(1), pp. 39-62.

Diggle, P.J., and M.G. Kenward (1994), “Informative drop-out in longitudinal data analysis”, *Applied Statistics*, 43, pp. 49–93.

Genz, A. (1992), “Numerical Computation of Multivariate Normal Probabilities”, *Journal of Computational and Graphical Statistics*, 1, pp. 141-149.

Genz, A. and K.S. Kwong (2000), “Numerical Evaluation of Singular Multivariate Normal Distributions”, *Journal of Statistical Computation and Simulation*, 68, pp. 1-21.

Hayter, A. J. (2006), “Recursive integration methodologies with statistical applications”, *Journal of Statistical Planning and Inference*, 136, pp. 2284 – 2296.

Heckman, J.J. and S. Navarro (2007), “Dynamic discrete choice and dynamic treatment effects”, *Journal of Econometrics*, 136(2), pp. 341-396.

Jacoby, S. L. S., Kowalik, J.S. and Pizzo, J.T. (1972), Iterative Methods for Nonlinear Optimization Problems, Englewood Cliffs, NJ: Prentice Hall, p. 81.

Kinderman, A.J. and J.G. Ramage (1976), ``Computer Generation of Normal Random Numbers,'' *Journal of the American Statistical Association*, 71, No. 356, pp. 893-896.

Lahiri, K., and J. Song (2000), "The effect of smoking on health using a sequential self-selection model", *Health Economics,* 9(6), pp. 491-511.

Lee, L.F. (1995), "Semiparametric maximum likelihood estimation of polychotomous and sequential choice models", *Journal of Econometrics*, 65(2), pp. 381-428.

Lu, M., and T.G. McGuire (2002), "The Productivity of Outpatient Treatment for Substance Abuse", *The Journal of Human Resources,* 37(2), pp. 309-335.

Manski, C.F. (1977), "The structure of random utility models," *Theory and Decision*, 8(3), pp. 229-254.

Manski, C.F. and E. Tamer (2002), *"*Inference on Regressions With Interval Data on a Regressor or Outcome*", Econometrica*, Vol. 70, No. 2 (March, 2002), 519–546.

Manski, C.F. (2004), "Social Learning from Private Experiences: The Dynamics of the Selection Problem", *Review of Economic Studies*, Vol. 71, No. 2 (Apr., 2004), pp. 443-458

Manski, C F. (2005), "Random Utility Models with Bounded Ambiguity", paper presented at conference in honor of Daniel McFadden, Berkeley, California, May 6-7, 2005.

McDonald, H.P., A.X. Garg and R.B. Haynes (2002), "Interventions to Enhance Patient Adherence to Medication Prescriptions", *Journal of American Medical Association*, 288(22), pp. 2868-2879.

McFadden, Daniel (1974), "Analysis of Qualitative Choice Behavior," in Paul Zarembka (ed.), Frontiers in Econometrics (New York: Academic Press).

Osterberg, L. and Blaschke, T. (2005), "Adherence to Medication", The New England Journal of Medicine, 353, pp. 487-97.)

Richard , J.F. and W Zhang (2007), "Efficient high-dimensional importance sampling", *Journal of Econometrics,* 141, pp. 1385–1411.

Rosen, A.B., A.B. Spaulding, D. Greenberg, J.A. Palmer, and P.J. Neumann (2009), "Patient , Adherence: A Blind Spot in Cost-Effectiveness Analyses?", *American Journal of Managed Care,* 15(9), pp. 626-632.

Sackett (1976),"Introduction" and "The Magnitude of Compliance and Noncompliance" in Sackett DL, Haynes RB, eds. Compliance With Therapeutic Regimens. Baltimore, MD: The Johns Hopkins University Press; 1976, pages 1-25.

Stuart, E.A., D.F. Perry, H.N. Le and N.S. Ialongo (2008), “Estimating Intervention Effects of Prevention Programs: Accounting for Noncompliance”, *Prevention Science,* 9(4), pp. 288-298.

Trost, R.F. and L.F. Lee (1984), “Technical Training and Earnings: A Polychotomous Choice Model With Selectivity”, *The Review of Economics and Statistics,* 66(1), pp. 151-156.

Table 1: Descriptive statistics of the variables used in the main model

Variable	Mean
Attendance	
Session 1	0.853
Session 2	0.819
Session 3	0.784
Session 4	0.710
Session 5	0.701
Session 6	0.674
Session 7	0.715
Demographics	
Male	0.264
Gender Not Reported	0.045
Excluded: Female	0.691
Black or African-American	0.019
Hispanic	0.328
Native American	0.019
Other races	0.015
Race Not Reported	0.051
Excluded: White	0.568
Age (Std Dev in parenthesis)	39.26 (8.36)

Note: All variables are binary indicators, except for the age variable.

Table 2: ML Estimates of B matrix

Variable	Session 1	Session 2	Session 3	Session 4	Session 5	Session 6	Session 7
Male	-0.019	0.028	-0.158**	0.011**	-0.107**	-0.043	-0.008
	(0.014)	(0.031)	(0.063)	(0.005)	(0.047)	(0.032)	(0.005)
Gender Not Reported	-0.084**	-0.019	0.466***	-1.007***	-0.285***	1.081***	0.551***
	(0.038)	(0.019)	(0.171)	(0.320)	(0.092)	(0.265)	(0.147)
Excluded: Female							
Black or African-American	-0.432	-0.447***	-0.364***	-0.123**	-0.494**	-1.149***	0.144
	(0.276)	(0.163)	(0.116)	(0.054)	(0.210)	(0.333)	(0.099)
Hispanic	-0.053*	0.067	0.258***	0.078**	0.305***	-0.019*	0.023**
	(0.031)	(0.044)	(0.067)	(0.034)	(0.065)	(0.011)	(0.010)
Native American	-0.018**	1.398***	-0.299*	-0.250**	-0.063***	0.092**	0.159*
	(0.008)	(0.309)	(0.160)	(0.123)	(0.023)	(0.045)	(0.082)
Other races	-0.448**	0.961**	-0.014	1.693**	0.899***	-0.470**	0.089***
	(0.193)	(0.464)	(0.023)	(0.726)	(0.313)	(0.204)	(0.032)
Race Not Reported	-0.761**	0.055*	-0.041	1.018**	0.934***	-0.050*	-0.346**
	(0.323)	(0.029)	(0.025)	(0.413)	(0.241)	(0.026)	(0.139)
Excluded: White							
Age	0.000	0.002	0.009***	0.012***	0.000	0.012***	-0.001
	(0.000)	(0.001)	(0.002)	(0.002)	(0.001)	(0.002)	(0.001)
Intercept	6.843***	1.772***	0.352***	-0.211**	-0.098	-0.720***	-0.124
	(1.076)	(0.413)	(0.070)	(0.093)	(0.114)	(0.102)	(0.108)

*** $p<0.01$; ** $p<0.05$; *$p<0.10$
Bootstrapped Std Dev in parenthesis.

Table 3: ML Estimates of $\mathbf{C}'$ Matrix

Random Error Term	Session 2	Session 3	Session 4	Session 5	Session 6	Session 7
Initial error, ε	1.978*** (0.367)	0.915*** (0.159)	0.646** (0.316)	0.483*** (0.145)	0.614*** (0.262)	0.943*** (0.214)
Session 1 error, ν_1	0.378 (0.269)	1.038*** (0.303)	0.470*** (0.176)	1.159*** (0.191)	0.658*** (0.192)	0.852*** (0.199)
Session 2 error, ν_2	n.a.	0.708*** (0.123)	1.856*** (0.646)	0.797** (0.348)	1.159* (0.228)	0.736*** (0.204)
Session 3 error, ν_3	n.a.	n.a.	1.415*** (0.216)	0.979*** (0.347)	0.622** (0.296)	1.160*** (0.272)
Session 4 error, ν_4	n.a.	n.a.	n.a.	0.855*** (0.129)	1.943*** (0.259)	1.268*** (0.224)
Session 5 error, ν_5	n.a.	n.a.	n.a.	n.a.	0.907*** (0.172)	1.546** (0.286)
Session 6 error, ν_6	n.a.	n.a.	n.a.	n.a.	n.a.	0.410 (0.261)

*** p<0.01; ** p<0.05; * <0.10

Bootstrapped Std Dev in parenthesis.