The Effects of Vocational Rehabilitation for People with Mental Illness

David Dean
University of Richmond

John Pepper
University of Virginia

Robert Schmidt
University of Richmond

Steven Stern*
University of Virginia

July 8, 2011

1 Introduction

The public-sector Vocational Rehabilitation (VR) program is a $3 billion federal-state partnership designed to provide employment-related assistance to persons with disabilities. While thought to play an important role in helping persons with disabilities to engage in gainful employment (Loprest, 2007), very little is known about the long term-efficacy of VR in the United States. With one very recent exception (Cimera, 2010), the last published evaluation of the U.S. public-sector VR program is from over 20 years ago (Dean and Dolan,1991).[1] Although certainly informative, the earlier studies have a number of methodological shortcomings and have only limited relevance to the current VR system which serves a clientele with a much wider range of impairments. Originally established in 1919 to provide restorative services to persons with primarily physical disabilities, the program's emphasis has shifted in recent decades to serve persons with cognitive impairments or mental illness. While comprising an ever-larger share of the VR clientele, the latter group has turned out to be particularly hard to serve. As the Government Accountability Office (2005)

*We would like to thank Joe Ashley, John Phelps, Kirsten Rowe, Ann Stanfield, Vlad Mednikov, and Jim Rothrock from Virginia Department of Rehabilition Services (DRS), David Stapleton, and other members of the National Institute of Disability and Rehabilitation Research (NIDRR) Advisory Group for excellent help and advice and Rachel Fowley for excellent research assistance. The DRS, the Virginia Department of Medical Assistive Services, the NIDRR, and the University of Virginia Bankard Fund for Political Economy provided generous financial support. All errors are ours.

[1] Also see Conley (1969); Bellante, (1972); Worrall (1978); Berkowitz (1988). Several more recent studies evaluate the European active labor market programs for persons with disabilities (e.g., Raum and Torp, 2001; Bratberg, Grasdal, and Risa, 2002; Frolich, Heshmati, and Lechner, 2004; and Aakvik, Heckman and Vytlacil, 2005).

notes, persons classified with mental or psycho-social impairments make up almost one-third of VR program exiters nationwide in 2003 but, at 30%, had the lowest employment rate outcome of all groups served. Consequently, an increasing share of VR expenditure, along with research and practice in the VR and mental health fields, has been concentrated on increasing the employability of persons with psychiatric disabilities.[2]

In this paper, we study the impact of the VR program using a unique panel data source on all persons who applied for services in the state of Virginia in State Fiscal Year 2000. We focus our attention on VR clients diagnosed with mental illnesses, an increasingly important part of the VR caseload. Kessler et al. (2001) estimates that more than 25% of U.S. adults had a mental illness in the previous year, with 7% having a major depressive disorder and 18% having anxiety disorders. The prevalence of mental illness among adults in the United States imposes severe employment consequences with unemployment rates for persons with severe mental illness estimated to be as high as 95% (Mueser, Salyers, and Mueser, 2001).

In addition to updating the existing evaluations of VR services in the United States, we make a number of notable contributions to the empirical literature. At the most basic level, we focus on the impact of VR services on clients with a specific type of impairment, mental illness, rather than the entire caseload. Except for Dean and Dolan (1991), the existing state-level evaluations of VR services distinguish among clients with mental illness, cognitive impairments, and physical impairments only by including dummy variables in regressions. The impact of VR services, however, is thought to differ by the type of limitation (Dean and Dolan, 1991; Baldwin, 1999; and Marcotte, Wilxox-Gok, and Redmond, 2000).

Importantly, our administrative data from the 2000 applicant cohort in Virginia allows us to make a number of contributions to the literature. Other economic analyses of VR efficacy (see Conley, 1969; Bellante, 1972; Worrall, 1978; Nowak, 1983) have relied almost exclusively on the Rehabilitation Service Administration's RSA-911 Case Service Report of nationwide closures from the VR program. The problems with evaluations based on these RSA-911 data are manifold. First, a censoring problem arises because the RSA-911 sample frame is drawn from cohorts of cases terminated from the program during the same year. This is a significant drawback for a program with a wide variation in program duration that results in comparing cohorts who applied for services

[2] The increased emphasis on achieving competitive employment outcomes for persons with psychiatric disorders has led to numerous studies published in the VR literature that examine specific interventions for persons with varying degrees of mental illness. See Bond, Drake, and Becker (2001) for a review of such analyses or Cook et al. (2005); Burns et al. (2007); or Campbell et al. (2010) for descriptions of specific experiments. These investigations typically consist of small clinical trials of a specific intervention of supported employment versus the more traditional VR practice of "train and place." Such randomized clinical studies are typically of short duration and thus lack sufficient information on longer-term employment outcomes. Ultimately, this type of analysis is not suited for evaluating the on-going VR program, which legally is not allowed to engage in randomized control studies using federal support.

over different time periods. By focusing on an applicant cohort, we avoid this censoring problem. Second, the RSA-911 reports earnings only at two points: 1) self-reported weekly earnings at the time of referral to the VR program and 2) following three months of employment if employed. As Loprest (2007) notes, these analyses suffered from the RSA-911's lack of longitudinal earnings. In our data, we observe quarterly employment and earnings data as well as VR service data from 1995 to 2008. Thus, using data on individual quarterly employment and earnings prior to, during, and after service receipt, we examine both the short- and long-term effects of VR services. Finally, evaluations using the RSA data classify clients as either receiving or not receiving substantial VR services. In practice, however, VR agencies provide a wide range of different services which are likely to have very different labor market effects. Using the administrative data from Virginia, we examine the impact of specific types of services rather than just a single treatment indicator. In particular, following Dean et al. (2002), we aggregate VR services into six types – diagnosis and evaluation, training, education, restoration, maintenance and other – and allow these six services to have different labor market effects.

Finally, we formalize and estimate a structural model of endogenous service provision and labor market outcomes. Except for controlling for observed covariates, the existing literature does not address the selection problem that arises if unobserved factors associated with VR service receipt are correlated with latent labor market outcomes. Hotz (1992) provides a framework for the Governmental Accountability Office that laid out several options for evaluation of the public-sector VR program in a non-experimental setting that included both parametric and non-parametric techniques to control for the problem of selection bias inherent in such voluntary programs. Although several studies of the European active labor market programs for persons with disabilities incorporated such methodologies (e.g., Raum and Torp, 2001; Frolich, Bratberg, Grasdal, and Risa, 2002; Heshmati, and Lechner, 2004; and Aakvik, Heckman, and Vytlacil, 2005), evaluations of VR programs in the U.S. have not kept up with the significant advances made during the past two decades in evaluations of manpower training programs (see, for example, Imbens and Wooldridge, 2009).[3] We address the selection problem using instrumental variables that are assumed to impact service receipt but not the latent labor market outcomes, pre-program labor market outcomes that control for differences between those who will and will not receive services, and a formal structural model of the selection process.

The paper proceeds as follows: Section 2 describes the economic model used throughout the paper. We construct a multivariate discrete choice model for service provision choices. We augment that with a probit-like employment equa-

[3] Dean and Dolan (1991) follow advances in the more general field of manpower training evaluation at the time (see, for example, Ashenfelter, 1978; Bassi, 1984; and Heckman and Hotz, 1989), but do not address the problem of selection on unobservables. Selection is thought to be a central problem in addressing the impact of job training programs (Card and Sullivan, 1988; LaLonde, 1995; Friedlander et al., 1997; and Imbens and Wooldridge, 2009). Aakvik, Heckman, and Vytlacil (2005) find that that this selection problem plays an important role in the evaluation of a Norwegian VR training program.

tion and an earnings equation. We allow for correlation of errors among all of the equations. Next, we describe the three sources of data used in our analysis in Section 3 and the econometric methodology used to estimate the model from Section 2 in Section 4. Estimation results are presented in Section 5, and a return on investment analysis is presented in Section 6. Our results imply generally high rates of return but with significant variation in returns across people with varying characteristics.

2 Model

Let y_{ij}^* be the value for individual i of participating in VR service j, $j = 1, 2, .., J$, and define $y_{ij} = 1\left(y_{ij}^* > 0\right)$ be an indicator for whether i receives service j. Assume that

$$\begin{aligned} y_{ij}^* &= X_i^y \beta_j + u_{ij}^y + \varepsilon_{ij}, \\ \varepsilon_{ij} &\sim Logistic \end{aligned} \tag{1}$$

where X_i^y is a vector of exogenous explanatory variables, and u_{ij}^y is an error whose structure is specified below. Next, let z_{it}^* be the value to i of working in quarter t, and define $z_{it} = 1\left(z_{it}^* > 0\right)$. Assume that

$$z_{it}^* = X_{it}^z \gamma + \sum_{k=1}^{K} d_{ik} \sum_{j=1}^{J} \alpha_{jk}^z y_{ij} + u_{it}^z + \upsilon_{it}^z \tag{2}$$

where X_{it}^z is a vector of (possibly) time-varying, exogenous explanatory variables, d_{ik} is a dummy variable equal to one iff the amount of time between the last quarter of service receipt and t is between time nodes τ_k and τ_{k+1},[4] and u_{it}^z is an error whose structure is specified below. The time periods implied by the nodes we use are a) 2 or more quarters before service, b) 1 quarter before service, c) 1 quarter after service to 8 quarters after service, and d) 9 or more quarters after service. Next let w_{it} be the log quarterly earnings of i at t, and assume that

$$w_{it} = X_{it}^w \delta + \sum_{k=1}^{K} d_{ik} \sum_{j=1}^{J} \alpha_{jk}^w y_{ij} + u_{it}^w + \upsilon_{it}^w \tag{3}$$

[4] In effect, we allow for level spline effects for service effects on labor market outcomes.

where variables are defined analogously to equation (2). Finally, assume that

$$\begin{aligned}
u_{ij}^{y} &= \lambda_{j1}^{y} e_{i1} + \lambda_{j2}^{y} e_{i2}, \\
u_{it}^{z} &= \lambda_{1}^{z} e_{i1} + \lambda_{2}^{z} e_{i2} + \eta_{it}^{z}, \\
u_{it}^{w} &= \lambda_{1}^{w} e_{i1} + \lambda_{2}^{w} e_{i2} + \eta_{it}^{w}, \\
\eta_{it}^{z} &= \rho_{\eta} \eta_{it-1}^{z} + \zeta_{it}^{z}, \\
\eta_{it}^{w} &= \rho_{\eta} \eta_{it-1}^{w} + \zeta_{it}^{w}, \\
\begin{pmatrix} \zeta_{it}^{z} \\ \zeta_{it}^{w} \end{pmatrix} &\sim iidN\left[0, \sigma_{\zeta}^{2} \begin{pmatrix} 1 & \rho_{\zeta} \\ \rho_{\zeta} & 1 \end{pmatrix}\right], \\
\begin{pmatrix} e_{i1} \\ e_{i2} \end{pmatrix} &\sim iidN\left[0, I\right], \\
\upsilon_{it}^{z} &\sim iidN\left[0, 1\right], \text{ and} \\
\upsilon_{it}^{w} &\sim iidN\left[0, \sigma_{w}^{2}\right].
\end{aligned} \tag{4}$$

We include the (e_{i1}, e_{i2}) to allow for two common factors affecting all dependent variables with factor loadings $\left(\lambda_{jk}^{y}, \lambda_{k}^{z}, \lambda_{k}^{w}\right)_{k=1}^{2}$. We also allow for serial correlation and contemporaneous correlation in the labor market errors $(\eta_{it}^{z}, \eta_{it}^{w})$. The covariance matrix implied by this error structure is presented in Appendix 8.1. See Dean et al. (2010a) for a similar structure applied to people with cognitive impairments.

3 Data

We use two main sources of data: a) the administrative records for the state fiscal year (SFY) 2000 applicant cohort of the Virginia Department of Rehabilitative Services (DRS), and b) the quarterly administrative records of the Virginia Employment Commission (VEC) from 1995 to 2008 for those people in the DRS data. We also merge these files with data from the Bureau of Economic Analysis on county-specific employment patterns. Each of these is discussed in turn below.

3.1 DRS Sample Frame

Our starting point is the administrative records of the Virginia DRS for the 10323 individuals who applied for vocational rehabilitative (VR) services in SFY 2000 (July 1, 1999 - June 30, 2000). Our analysis focuses on 1555 DRS clients with mental illnesses. We exclude individuals for the reasons specified in Table 1. The criterion associated with having a mental illness used for sample selection is that the primary or secondary diagnosis listed in the administrative records must be a mental illness in at least one quarter while the individual has an open case; this may be the first case in 2000, or it may be a subsequent case. Not having a mental illness is the single most important reason for exclusion from our estimation sample, resulting in 6476 excluded observations. Because we

Table 1: Missing Value Analysis

Cause	# Obs Lost	Proportion of Total	# Remaining
Applicants in SFY 2000			10323
Missing or Questionable SSN	81	0.008	10242
Died While in Program	65	0.006	10177
Missing Gender or Date of Birth	1	0.000	10176
Not in Virginia	59	0.006	10117
Not Mentally Ill	6476	0.640	3641
Missing Primary Disability	87	0.024	3554
Missing Secondary Disability	7	0.002	3547
Initial Service Spell before SFY 2000	1220	0.344	2327
Age Younger than 21 Years	701	0.301	1626
Neither Service nor Employment Record	71	0.044	1555
Number Remaining in Sample	1555	0.151	

need diagnoses for each case, we exclude 94 observations where primary and/or secondary diagnosis was missing as well.[5] We also excluded 71 individuals with neither any service records nor employment records.[6]

We focus on the "base case" defined as an individual's initial case in SFY 2000, recognizing that individuals can have multiple "service spells" or "cases," each of which includes an application and administrative closure. We have administrative information between SFY 1987 and 2007 that allows us to identify these multiple service spells and exclude observations where the individual's first service spell was prior to SFY 2000. We do this to avoid bias associated with left censoring (e.g., Heckman and Singer, 1984a). In particular, if the subsample of people who enroll in services more than once is different than those who enroll only once, then those people who had service spells prior to SFY 2000 will have unobservable characteristics different than those whose first spell is in SFY 2000.

3.2 DRS Data for Service Provision

Upon application, an individual's case is assigned to a counselor who assesses the individual's eligibility for the program. This assessment typically includes a diagnosis of the impairment. The case may be administratively closed at this point because the impairment is deemed insufficiently severe or too severe or

[5] All of the occurances of missing primary and/or secondary diagnosis occur in subsequent cases after the SFY 2000 base case.

[6] While it could be the case that such individuals applied to DRS and withdrew for some reason and were also never employed, we were concerned about including such observations because there was a reasonable chance of a problem with the merging of the DRS and VEC data. To the degree that we excluded valid observations, we are biasing our results toward finding no effect for DRS services because the excluded observations would have been recorded as having no employment and no change in employment before and after service had we included them.

because the individual withdraws from further consideration for VR eligibility. Beyond assessment and some counseling, these individuals receive few, if any, services.

By contrast, for those accepted for service, the counselor and individual develop an individualized plan for employment (IPE) which specifies the array of services to be provided. Services can include, for example, restorative medical care, vocational counseling and guidance, training (both vocational and rehabilitative), education, job search and placement, and/or assistive services. Some individuals drop out before completing the program, possibly having received little or no services beyond the development of an IPE.

Services can be provided to an individual in any combination of three ways: a) internally by DRS personnel, b) as a "similar benefit" purchased or provided by another governmental agency or not-for-profit organization with no charge to DRS, and/or c) as a "purchased service" through an outside vendor using DRS funds. We have access to dates, quantities, costs, and types of purchased services because they are recorded by DRS for accounting purposes. Purchased service expenditures were recorded for 70% of base cases, ranging up to \$48069 for a single case with a mean of \$2055 and standard deviation of \$3848. However, we do not observe the same information for either in-house services or similar benefits. DRS records their provision (but not timing or cost) for the Rehabilitation Service Administration RSA-911 Case Service Report due at the end of the federal fiscal year for all cases closed during that year. The use of this information is limited further by a change in federal reporting standards during the period in which these cases were active. Many of the cases had closed before the new standards were implemented.[7]

Given these data limitations, our model focuses exclusively on the receipt of purchased services, although we do impute a value of total service costs for cost/benefit analysis as discussed in Section 6. We ignore service cost in the estimation model because of a) incomplete service costs and provision dates and b) the standard approach for evaluating labor market training and VR programs is to focus on binary indicators of service provision (see, for example, Dean and Dolan, 1991; LaLonde, 1995; Friedlander, Greenberg, and Robins, 1997; Heckman, LaLonde, and Smith, 1999; Imbens and Wooldridge, 2009).

There are 76 separate services provided by DRS, other state agencies, and 1252 vendors.[8] Following Dean et al. (2002), we aggregate these services into the six service types listed in Table 2. As discussed above, *diagnosis & evaluation*[9] are provided at intake in assessing eligibility and developing an IPE and possibly later in the form of job counseling and placement services. *Training* includes

[7] We examined the seriousness of this issue for an earlier cohort that had detailed federal codes. For that disability type, the majority of in-house services and similar benefits appear to have been provided during the application process in the form of assessment and the development of an individual employment plan rather than for program services (Dean et al., 2010a). The estimated impact of diagnostic services should be qualified accordingly.

[8] Of the 1252 vendors, 73 are employment service organizations which receive roughly half of total purchased-service dollars, usually in the form of job coach services or supported employment.

[9] We put variable names in a different font to avoid confusion.

Table 2: Proportion Receiving DRS Purchased Service by Type and Spell

Variable	Initial Service Spell	After Initial Service
# Observations	1555 individuals	49012 quarters
Diagnosis & Evaluation	0.490	0.015
Training	0.322	0.030
Education	0.111	0.008
Restoration	0.286	0.015
Maintenance	0.257	0.022
Other Service	0.199	0.016

vocationally-oriented expenditures including those for on-the-job training, job coach training, work adjustment, and supported employment. *Education* includes tuition and fees for a GED (graduate equivalency degree) program, a vocational or business school, a community college, or a university. *Restoration* covers a wide variety of medical expenditures including dental services, hearing/speech services, eyeglasses and contact lenses, drug and alcohol treatments, psychological services, surgical procedures, hospitalization, prosthetic devices, and other assistive devices. *Maintenance* includes cash payments to facilitate everyday living and covers such items as transportation, clothing, motor vehicle and/or home modifications, and services to family members. *Other services* consists of payments outside of the previous categories such as for tools and equipment.

Diagnostic and evaluation services are purchased in 49% of the base cases. Purchased services are provided in less than a third of the cases for every other service type. This should be qualified by noting that 16% of applicants are not accepted into the program, and another 30% drop out after acceptance but before receiving substantive services. Of the remaining applicants, 80% are provided a purchased service other than for *diagnosis & evaluation*.

The second column of Table 2 assesses the prevalence of subsequent service spells during the period in which we estimate employment impacts of service receipt. A total of 364 individuals (about 23% of the sample) returned after their base case for an additional 549 service spells, or an average of 1.5 additional cases apiece.[10] For the full 1555 individual sample, the second column shows the proportion receiving each service, disaggregated by service type, in any given quarter after the base-case closure date. Out of the 49012 person-quarters that we observe after closure, we observe receipt of training in 3.0% of them. In this paper, we ignore subsequent service spells in the structural model described in Section 2 in order to avoid having to model entry into and exit from services.

[10] Of these 364 individuals, 74 returned for a third service spell, 11 for a fourth, and 1 for a fifth.

Table 3: Frequency and Proportion of Most Common Service Combinations

Initial Service Spell			Subsequent Service Quarters		
Combination	Frequency	Proportion	Combination	Frequency	Proportion
d	232	0.18	t	670	0.20
dr	119	0.09	m	397	0.12
r	73	0.06	o	336	0.10
dt	63	0.05	d	297	0.09
t	52	0.04	r	255	0.08
tm	45	0.04	tm	194	0.06
dtm	43	0.03	dr	122	0.04
dtrmo	40	0.03	e	117	0.03
m	36	0.03	to	116	0.03
dtmo	31	0.02	dt	74	0.02
tmo	30	0.02	te	72	0.02
dtrm	29	0.02	rm	62	0.02
dtro	28	0.02	tmo	57	0.02
All others	439	0.35	All others	590	0.18

Notes:

1.d=diagnosis & evaluation, t=training, e=education, r=restoration, m=maintenance, and o=other service. Strings of letters imply receipt of each service in the string.
2.For initial service, the unit of observation is a person, and, for subsequent service, the unit of observation is a person-quarter.

We leave such analysis for future work.

As is seen in Table 3, a high proportion of clients receive multiple services during the same service spell. For example, while the most common service combination in the initial service spell is *diagnosis & evaluation* with no other service (d), the next most common is *diagnosis & evaluation* along with *restoration* (dr), and *diagnosis & evaluation* along with *training* (dt) is the fourth most common. Given the frequency with which clients receive multiple services, it is critical for us to allow for the possibility of receipt of multiple services. Thus, the structure of the service choice in equation (1) is multivariate discrete choice rather than polychotomous discrete choice.

Throughout much of the analysis, we measure labor market outcomes relative to the initial service period, defined as the first quarter in which purchased services are provided.[11] While this is a simple and appealing way to define the

[11] We construct the initial service period as lasting one quarter and with the quarter of service t_s defined in the following way: Define a_s as the starting quarter of the initial case and a_e as the ending quarter of the initial case. Let $s_{it} = 1$ iff i received purchased service in quarter t for $a_s \leq t \leq a_e$. Then

$$t_s = \begin{cases} \min\limits_{a_s \leq t \leq a_e} t : s_{it} = 1 & \text{if } \sum_{a_s \leq t \leq a_e} s_{it} > 0 \\ a_s & \text{if } \sum_{a_s \leq t \leq a_e} s_{it} = 0 \end{cases};$$

i.e., the quarter of service is the first quarter during the initial spell when purchased service is received (except when no service is received). In Figure 1, the curve labeled "Case Open vs Assumed Service Date" is the density of $\Delta_s = t_s - a_s$ and the curve labeled "Assumed date of service receipt, there are two potential shortcomings of this measure: first, the initial service quarter may differ from the application quarter, and second, some clients receive services over multiple quarters. Figure 1 provides information about the importance of these issues, with the curve labeled "Case Open vs Assumed Service Date" revealing the density of how long it takes (in quarters) to start receiving service after the application quarter, and the curve labeled "Assumed Service Date vs Last Service Receipt" displaying the density of the length of service receipt.[12] The first issue associated with the difference between the application and service dates is that one might want to treat labor market outcomes differently before and after application quarter (e.g., the Ashenfelter dip). Instead, we focus on a one-quarter pre-service dip in our specification of the model (see Section 2).The figure shows that 44% start receiving services in the application quarter and 83% start within 2 quarters. Meanwhile, 3% of DRS clients receive initial services 12 or more quarters after the application date. Thus, this issue may not matter that much given the concentration near zero. The second issue associated with the length of spells is that there may be a significant difference in labor market outcomes while service is being received and after it is finished. In our specification of the model, we distinguish between outcomes 8 or fewer quarters after service and 9 or more quarters after service. Figure 1 shows that 56.1% receive services for 3 quarters or less and only 19.1% of applicants are still receiving service after 8 quarters. Thus, for the most part, one can interpret the results for 9 or more quarters as being post-service receipt.

One alternative way to define post-service outcomes would be to use the closing date of the service spell as the end of service. This is the case for most of the literature (e.g., Dean and Dolan, 1991). The problem with this approach is that counselors do not close cases necessarily when service provision ends. Another way is to model the transition associated with the end of purchased service receipt. We think this is an important long-term research goal but beyond the goals of this paper. Alternatively, one could just use the end of service receipt as the quarter defining the beginning of relevant labor market outcomes; we were somewhat concerned with endogeneity issues associated with the length of service receipt and later labor market outcomes. While our approach has issues associated with it as well, its simplicity makes it a good place to start exploration of the data.

3.3 DRS Data for Explanatory Variables

Table 4 provides the sample moments for the explanatory variables coming from the DRS data to be used in the analysis. While many of the variables

Service Date vs Last Service Receipt" is the density of

$$\Delta_e = \left[\max_{a_s \leq t \leq a_e} t : s_{it} = 1\right] - t_s.$$

[12] The up-tick at 20 quarters occurs because of censoring imposed by us at 20 quarters for this figure.

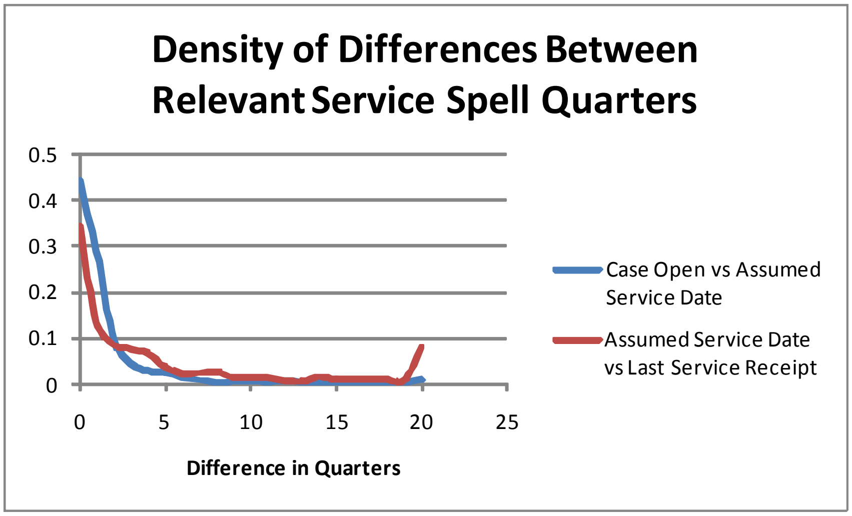

Figure 1: Density of Differences between Relevant Service Spell Quarters

are standard for this type of analysis, some are unusual and included because of the nature of the people being considered. *Special education* is a dummy variable equal to 1 for those observations where the respondent received some type of special education; 2.5% of the respondents received such education. Education information is missing for 10.3% of the sample. Rather than exclude such observations, instead we included a dummy variable for when education information was missing.[13]

There are a number of indicators of physical and mental disabilities in the DRS data. We use four dummy variables, each equal to one if the individual's primary or secondary disability at intake in the base SFY 2000 case was diagnosed as a musculoskeletal impairment, a learning disability, a mental illness, and a substance abuse problem.[14] An individual's counselor also assesses the significance of the disability resulting in substantial functional limitations and/or requiring multiple VR services over an extended period of time. Three levels are identified: *not significant* (used as the base level), *significant*, and *most significant*. We also constructed a dummy for serious mental illness (SMI) based on detailed diagnostic codes.[15]

[13] While this is a common way to address missing explanatory variables, Abbrevaya and Donald (2011) raises some concerns about using this approach.

[14] The existence of visual, hearing/speech, internal disabilities, and other miscellaneous disabilities and cognitive impairments were available in the data but not common enough or not varying enough with dependent variables to measure precise effects. So they were not used in the analysis.

[15] An individual was labeled as having a serious mental illness iff he/she had schizophrenia and/or psychosis. Given the information we have, this implies that the group of individuals with non-serious mental illness includes those with anxiety disorder, depression, personality disorder, and generic mental illness; substance abuse is measured separately.

Table 4: Moments of Explanatory Variables

Socio-Demographic Variables			Disability Variables		
Variable	Mean	Std Dev	Variable	Mean	Std Dev
Male	0.404	0.491	Type		
White	0.710	0.454	Musculoskeletal Disability	0.170	0.376
Education	10.718	4.931	Learning Disability	0.046	0.209
Special Education	0.025	0.156	Mental Illness	0.950	0.218
Education Missing	0.103	0.304	Substance Abuse Problem	0.151	0.358
Age (Quarters/100)	1.427	0.407			
Married	0.178	0.383	Extent		
# Dependents	0.804	1.171	Significant	0.619	0.486
Transportation Available	0.741	0.438	Most Significant	0.275	0.446
Has Driving License	0.678	0.467	SMI	0.236	0.425
Receives Govt Assistance	0.191	0.300			

While some variables such as *married* and *# dependents* may be endogenous, we follow the literature (e.g., Keith, Regier, and Rae, 1991; Ettner, Frank, and Kessler, 1997) and include them anyway as significant indicators of inclusion in society and responsibility. We include a dummy for receipt of government financial assistance even though it may be endogenous. However, for this population, one can work without losing one's government assistance or having it reduced up to relatively high earnings thresholds. Finally, we include two transportation variables: *transportation available* and *has driver's license*. Raphael and Rice (2002) worries about the endogeneity of these variables and finds that controlling for endogeneity with some reasonable instruments has little effect on the estimated effect of transportation on employment but makes its effect on wages disappear.

To identify the impact of services on labor market outcomes, we exploit two instrumental variables that are correlated with the treatment assignment but not included in the labor market equations (2) and (3). These instruments are the proportion of other clients in our cohort for the individual's counselor receiving a particular service and the proportion of other clients in our cohort for the individual's field office receiving a particular service. These variables are transformed as is described in the Appendix 8.2. The properties of these instruments depend upon the distribution of client size in our sample across counselors and field offices[16] and the distribution of the proportion of clients receiving each service. Figures 2, 3, and 4 provide some information about these distributions. In Figure 2, we see that there is significant variation in the size of counselor caseloads and field office caseloads. For example, 43% of counselors have caseloads from our cohort of 5 or less, and 7.3% have caseloads of 20 or more. Analogously, 36.7% of field offices have caseloads from our cohort of 10 or less, and 20.5% have caseloads of 50 or more.

Figure 3 shows the empirical distribution of proportion of clients for each field office receiving each service. For example, for *diagnosis & evaluation*,

[16] For example, if most counselors and/or field offices had only one client, then this methodology would not be useful.

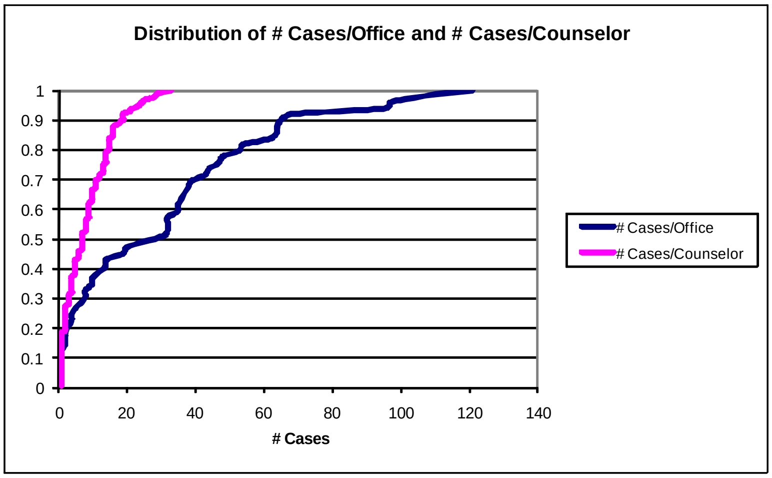

Figure 2: Distribution of # Cases/Office and # Cases/Counselor

10.4% provide the service to 18.2% of their clients or less, and 4.2% provide it for all of their clients. Figure 3 shows that *diagnosis* & *evaluation* is the most commonly provided service, followed by *training*, then *restoration* and *maintenance*, then *other services*, and then *education*. In fact, except for *restoration* and *maintenance* and some choices at very low levels of provision, each curve stochastically dominates the ones behind it across offices. Figure 4 has similar properties for counselors.

There is strong evidence of important variation in behavior across counselors and across field offices. We reject the null hypothesis that the joint density of services within offices does not vary across offices using a likelihood ratio test. The test statistic is 407.44 (with 245 df and normalized value of 7.33).[17] We also can test the null hypothesis that each office provides each service in the same proportion, one at a time, using a likelihood ratio test. The test statistic is 575.39 (with 294 df and a normalized value of 11.60). For counselors, the analogous test statistics are 970.60 (with 785 df and a normalized value of 4.68) and 3836.94 (with 942 df and a normalized value of 66.70). The fact that there is significant variation in the provision of services across offices and counselors make our instrument viable.

Table 5 shows that, while there is significant positive correlation across counselor and office effects, there is enough independent variation between them to accurately estimate their effects on service provision.

[17] For χ_k^2 random variables with large degrees of freedom k, one can normalize by subtracting the mean k and dividing by the standard deviation $\sqrt{2k}$; i.e.,

$$\frac{\chi_k^2 - k}{\sqrt{2k}} \sim N(0,1)$$

under H_0.

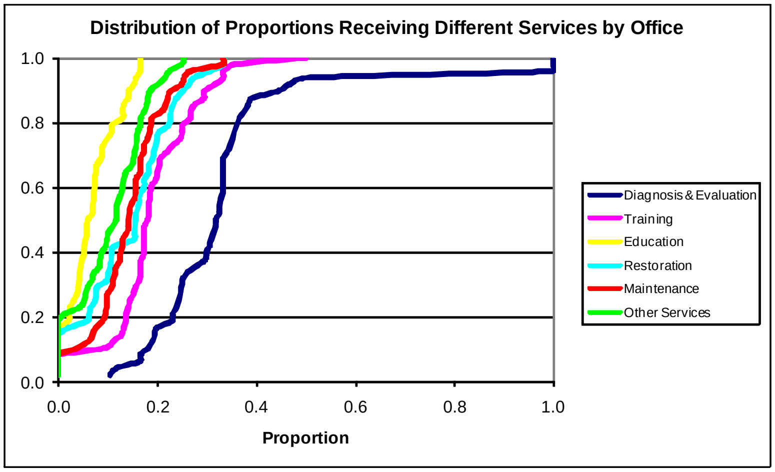

Figure 3: Distribution of Proportions Receiving Different Services by Office

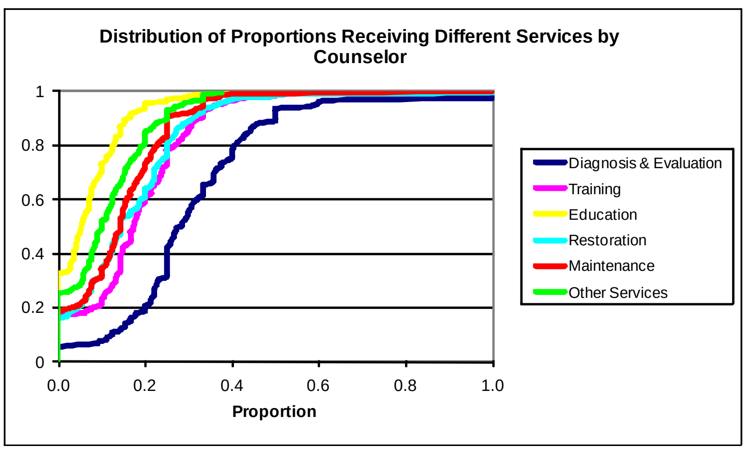

Figure 4: Distribution of Proportions Receiving Different Services by Counselor

Table 5: Correlation of Office and Counselor Variables

Service	Correlation	Service	Correlation
Diagnosis & Evaluation	0.592	Restoration	0.674
Training	0.455	Maintenance	0.632
Education	0.588	Other Service	0.710

3.4 VEC Data

One of the unique and valuable features of this analysis is that we have information from an administrative data source about individual quarterly earnings prior to, during, and after service receipt. Earlier economic analyses of VR efficacy (Conley, 1969; Bellante, 1972; Worrall, 1978; Nowak, 1983) relied almost exclusively on the RSA-911 Case Service Report of nationwide closures from the VR program. At the time, the 911 form reported earnings only at two points: 1) self-reported weekly earnings at the time of referral to the VR program and 2) following two months of employment. The latter figure is available only for that portion of VR cases closed "with an employment outcome." More recent analyses, published almost entirely in the rehabilitation literature (e.g., Cimera, 2010), utilize the same RSA-911 earnings measure, albeit now collected after three months of employment. In contrast, this study uses data gleaned from quarterly employment records provided by employers to the Virginia Employment Commission (VEC) for purposes of determining eligibility for unemployment insurance benefits.

The DRS provided the VEC with identifiers from the universe of 10323 applicants for DRS services in SFY 2000. The VEC returned to DRS a longitudinal file containing employment data for 9041 individuals having at least one quarter of "covered" employment during the 47-quarter period spanning July 1995 through March 2009, a "hit rate" of 88%. The remaining 12% in this cohort were either a) unemployed or out of the labor force for this entire interval or b) employed in jobs that are not covered by the VEC (e.g., were self-employed or worked out of state, for federal employers, for very small-sized firms, or at contingent-type jobs that do not provide benefits).

We explored the coverage issue through an arrangement with the Social Security Administration (SSA) whereby they matched VEC earnings (aggregated to a calendar year) to calendar-year SSA earnings for all SFY 2000 applicants.[18] Table 6 summarizes these results for the 9913 individuals with an identification match. For the two calendar years following SFY 2000 (the fiscal year of application), the SSA and VEC agreed on employment status for 87% of individuals. VEC records missed employment covered by SSA for 12% of the individuals in both 2001 and 2002. For those individuals where both SSA and VEC report earnings, VEC earnings levels fall short of SSA levels by 5.6% in 2001 and 6.2% in 2002.[19]

[18] This analysis was not limited to applicants with mental illness diagnoses.

[19] Data from the National Health Interview Survey 2004 Adult Sample (NHIS) shows that, for the United States as a whole, people with mental illness have probabilities of working for the federal government and being self-employed of 2.7% and 7.8%, respectively; corresponding numbers for those without mental illness are 3.0% and 8.4%, respectively. However, because of its proximity to Washington, DC and its large number of military facilities, Virginia has an unusually high proportion of federal workers; using data from the Bureau of Economic Analysis (2010b), the proportion of employed individuals in Virginia working for the federal government (including the military) in 2000 was 7.6%, while the NHIS data implies that it was 3.3% for the United States in 2004. If we conclude that $7.8\% + (7.6/3.3) * 2.7\% = 14.2\%$ of Virginians with mental illness either work for the federal government or are self-employed, this accounts for all of the discrepancy between SSA earnings and VEC earnings in the 2nd

Table 6: Comparison between SSA and VEC Employment Records

	2001	2002
Neither SSA nor VEC show earnings	31%	35%
SSA shows earnings, VEC does not	12%	12%
VEC shows earnings, SSA does not	1%	1%
Both SSA & VEC show earnings	57%	52%
Mean SSA Earnings	\$9,117	\$9,859
Mean SSA - VEC Difference	\$510	\$616

Employers report aggregate earnings in a given quarter to the VEC. Recall that equations (2) and (3) model employment and earnings impacts in four separate periods offset from the date of first service. Because the date of first service can fall anywhere within a quarter, that quarter is excluded from the analysis other than for use as a period of demarcation separating pre-service from post-service periods. Depending upon the date of first service, this alignment procedure results in 16 to 19 quarters of pre-service earnings periods and 28 to 31 quarters post-service quarters for individuals in this cohort.

In our analysis, we try to explain two labor market outcome variables: *employment* and *log quarterly earnings*. *Employment* is a binary measure of working in a particular quarter in the labor market and is modeled in equation (2). We also measure *log quarterly earnings* in equation (3). While it would be valuable to be able to decompose quarterly earnings into wage level and hours, this is not possible in the VEC data. Table 7 provides information on sample sizes and on the moments of employment data and earnings data disaggregated between quarters before and after initial service provision. The sample sizes are quite large and allow us to estimate labor market outcome effects with high precision. One can see that employment rates decline after service provision and quarterly earnings increase (conditional on working). However, as is shown in Section 5, these aggregate facts hide what is really happening and how it depends on service receipt.

Figures 5 and 6 display quarterly employment rates and earnings (conditional on employment), respectively, for SFY 2000 applicants who receive substantive VR services and those that do not receive substantial services. We refer to these two groups as the treated and untreated, respectively. In these figures, quarters are measured relative to application date (not the initial service date) so that quarter 0 is the quarter of application, quarter -4 is one year prior to application, and quarter 4 is one year post-application.

row of Table 6.

Table 7: Moments of Employment and Earnings Variables

	Before Initial Service Quarter			After Initial Service Quarter		
Variable	# Obs	Mean	Std Dev	# Obs	Mean	Std Dev
Employment	31427	0.35	0.477	58763	0.275	0.446
Log Quarterly Earnings	11003	7.082	1.492	16145	7.519	1.398

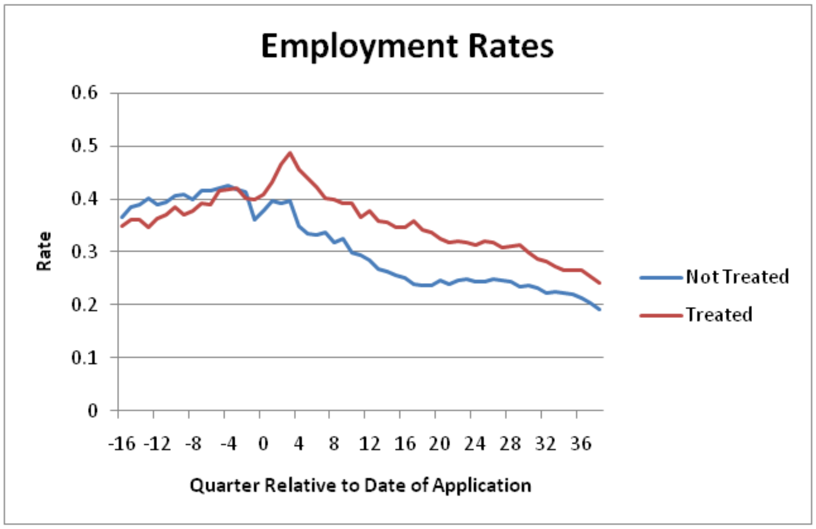

Figure 5: Employment Rates

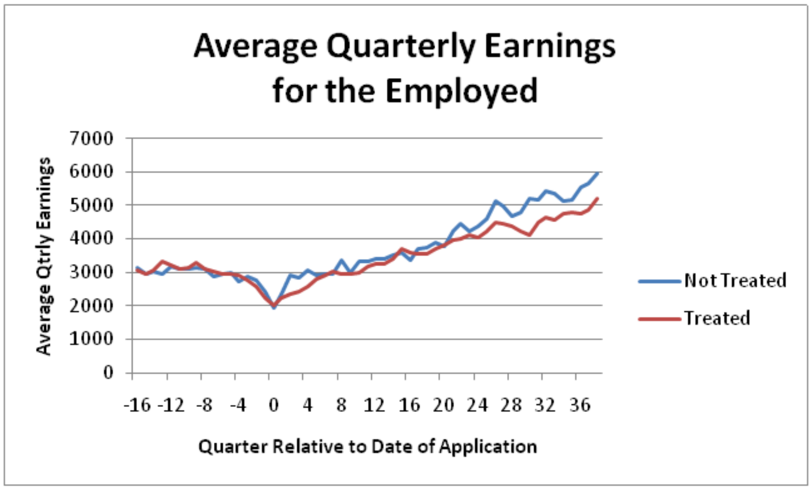

Figure 6: Average Quarterly Earnings for the Employed

Perhaps the most striking finding is seen in Figure 5 which shows that, prior to the application quarter, the employment rates of the treated and untreated are nearly identical, with a modest Ashenfelter dip in the pre-application quarter, but, just after the application quarter, the treated experience a pronounced increase in employment rates. For example, one year prior to the application quarter, the employment rates are 0.42 for both the untreated and treated, while, one year after the application, the analogous employment rates are 0.35 for the untreated and 0.46 for the treated. About one year after the application, the employment rates for both the treated and untreated start to decline, but a gap continues between the two groups. After nine years, the employment rates of around 0.20 are notably less than the rates in SFY 2000.

While there is notable association between DRS services receipt and employment, there is no such relationship with earnings. Figure 6 shows that quarterly earnings among the employed are almost identical for the treated and the untreated throughout. Thus, the data reveal that VR treatment services are associated with a sharp, substantial, and sustained increase in employment but no discernible change in quarterly earnings among the employed.

Figures 5 and 6 also shed some light on the appropriate assumption about the length of the Ashenfelter dip. Depending on the program being evaluated, the pre-program dip in employment and earnings has been generally found to start between one quarter and one year prior to participation in the program (Heckman et al., 1999; Mueser et al., 2007). For our sample, Figures 5 and 6 reveal a dip in earnings in the first quarter prior to the initial service receipt. Thus, we account for the Ashenfelter dip using a one quarter pre-service indicator in employment and earnings equations.

Table 8 provides some information on the frequency of transitions, starting in the application period, in the merged DRS/VEC data.[20] For example, we observe 551 client-quarters where clients moved from receiving services in their initial service spell to becoming employed in the next quarter.[21] As one would expect, the diagonal elements of the transition matrix are dominant, implying that service spells, employment spells, and non-employment spells are likely to last more than one quarter. However, we also observe much movement in the sample between service receipt and the labor market and between employment and non-employment. In particular, the probability of transitioning from employment to non-employment is about 13% ($= 2370/(15222 + 253 + 217)$), and the probability of going from non-employment to employment is 5%. Similar patterns of state dependence are found in the general population, but VR clients have a much weaker attachment to employment; there is a much higher probably of exiting employment and a much smaller probability of transitioning out of non-employment to employment. Shimer (2007), for example, reports employment exit probabilities that average about 3.5% and job finding probabilities of

[20] For the purposes of analyzing transitions in this section, a service spell ends when there is a period with no receipt of any services. For the remainder of the paper, only the first service spell is used, and it ends in one quarter.

[21] We cannot distinguish between unemployment and "not-in-the-labor-force." Thus, we call all spells not working and not receiving services "non-employment."

Table 8: Transition Matrices in Frequencies

	Initial Service	Employed	Not Employed	Subsequent Service
Initial Service	1841	551	709	0
Employed	217	15222	2370	253
Not Employed	375	1954	36978	390
Subsequent Service	0	303	340	874

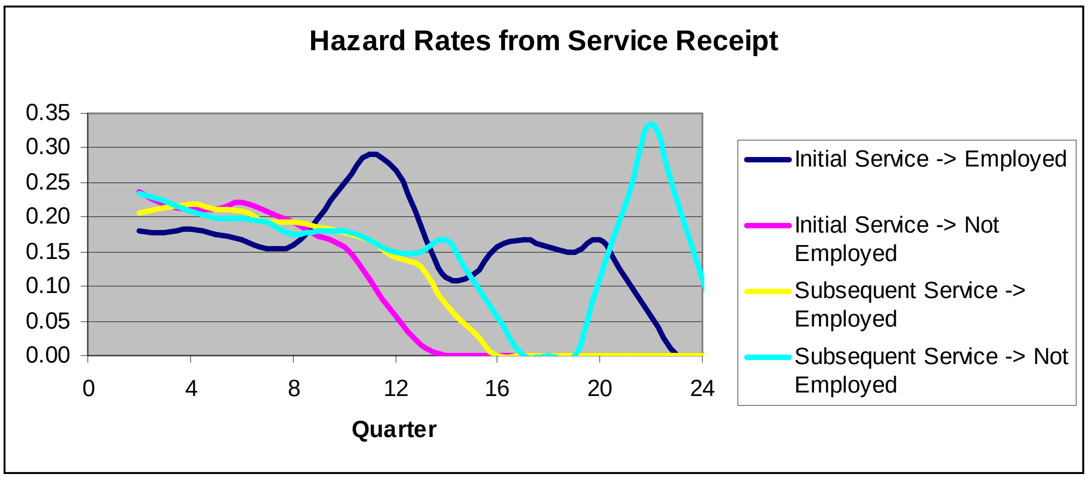

Figure 7: Hazard Rates from Service Receipt

about 45% (also, see Gangl, 2003; Hall, 2005; and Farber, 2003).

The transition matrix also shows the frequency of recidivism back to DRS for more services. Among employed people, the return rate is 1.4% ($= 253/(15222 + 2370 + 253)$), and, among non-employed people, it is 1.0% ($= 390/(1954 + 36978 + 390)$). For this paper, we do not model the 1% of transitions from the labor market to subsequent service spells.

One of the implications of Table 8 is the inertia associated with VR service receipt, employment, and non-employment spells. Figures 7 and 8 provide information on hazard rates across the transitions associated with Table 8. The reported empirical hazards out of service receipt in Figure 7 show that, while many clients leave service to start a job, a significant number leave without a job. For the first two years of service receipt, initial service spells and subsequent service spells have similar hazard rates. However, after two years, success into jobs is better for people in their initial service spell.

Figure 8 shows hazard rates out of employment and out of non-employment. Here, as in Table 8, we see that transitions between employment and unemploy-

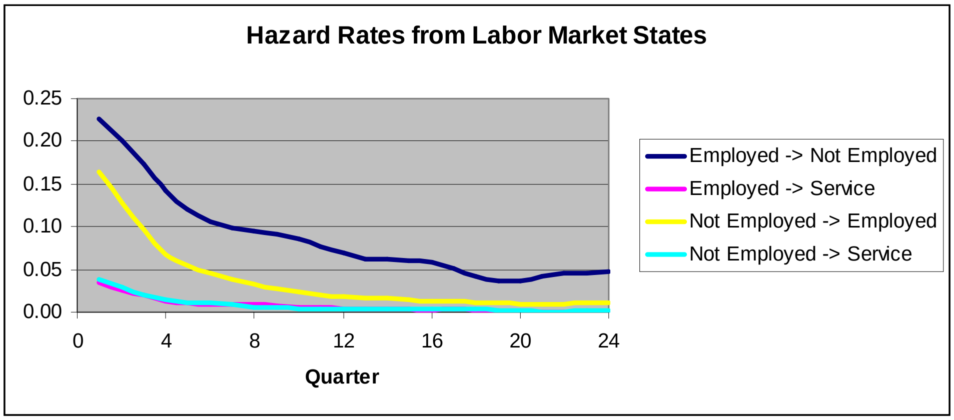

Figure 8: Hazard Rates from Labor Market States

ment are much higher than into service provision. Moreover, the probability of leaving employment to unemployment (or vice versa) is higher in the first few quarters suggesting relatively frequent transitions between labor market states. After a few quarters, however, the probability of exiting a particular labor market state sharply drops, and the survival function becomes nearly flat. Thus, the transition probabilities are highly sensitive to the length of time in a particular state. This pattern of negative duration dependence is consistent with other work on larger populations in the labor economics literature (e.g., Moffitt, 1985; Van Den Berg and Van Ours, 1996; Machin and Manning, 1999; and Bover, Arellano, and Bentolila, 2002). Dean et al. (2010a) also find similar transition patterns among a sample of DRS clients with cognitive impairments.[22] Declining duration dependence may occur for two reasons. First, as an individual gains experience in a particular job he gains skills that are specific to that job, making him more valuable at that job than others and reducing turnover. Second, the workers who leave a job early are, on average, not good fits for that job, leaving the remaining workers as better fits on average. In our model, described in Section 2, we do not allow for inertia as is seen to exist in Figures 7 and 8. While modelling such inertia using a flexible hazard rate specification (e.g., Heckman and Singer, 1984b; and Meyer, 1990) would be useful, it is not that important for our goal in estimating rates of return. So we leave it for future research.

3.5 BEA Data

Labor market outcomes may be influenced by local labor market conditions. Though there are no measures of local labor market conditions in either the

[22] Butler, Johnson, and Baldwin (1995) state that evaluations of worker injuries that rely on the single episode model will overstate the success of interventions, such as medical care or rehabilitative services, by treating the first return to work as a permanent, stable outcome.

Table 9: Moments for Local Labor Market log Employment Rate Variables

Geography	# Obs	Mean	Std Dev
County	91072	-0.505	0.230
MSA/RSA	91072	-0.500	0.229

DRS data or the VEC data, the DRS data contain geographic identifiers so that we can match each DRS client with their county of residence. The Bureau of Economic Analysis (BEA) provides information on population size and number of people employed, disaggregated by age and county (BEA, 2010a). We construct measures of log employment rates using two units of geography: county and MSA/RSA level. Details are included in Appendix 8.3. Moments are provided in Table 9. The correlation between the two measures is 0.99; however, given the large sample size, the inner product of explanatory variables is not singular.

4 Econometric Methodology

4.1 Likelihood Function

The parameters of the model are $\theta = (\theta_y, \theta_z, \theta_w)$ where

$$\begin{aligned}
\theta_y &= \left(\beta_j, \lambda_{j1}^y, \lambda_{j2}^y\right)_{j=1}^J, \\
\theta_z &= \left(\gamma, \lambda_1^z, \lambda_2^z, \rho_\eta, \sigma_\zeta^2, \rho_\zeta, \left[\alpha_{jk}^z\right]_{j=1}^J\right), \text{ and} \\
\theta_w &= \left(\delta, \lambda_1^w, \lambda_2^w, \rho_\eta^w, \sigma_w^2, \left[\alpha_{jk}^w\right]_{j=1}^J\right).
\end{aligned}$$

We estimate the parameters of the model using maximum simulated likelihood (MSL). The likelihood contribution for observation i is

$$L_i = \int L_i(u_i)\, dG(u_i \mid \Omega) \tag{5}$$

where

$$L_i(u_i) = L_i^y(u_i^y) \prod_{t=1}^T L_{it}^{zw}(u_{it}^z, u_{it}^w),$$

$$L_i^y(u_i^y) = \prod_{j=1}^J \frac{\exp\left\{X_i^y \beta_j + u_{ij}^y\right\}}{1 + \exp\left\{X_i^y \beta_j + u_{ij}^y\right\}}, \tag{6}$$

$$L_{it}^{zw}(u_{it}^z, u_{it}^w) = \left[L_{it}^0(u_{it}^z, u_{it}^w)\right]^{1-z_{it}} \left[L_{it}^1(u_{it}^z, u_{it}^w)\right]^{z_{it}}, \tag{7}$$

$$L_{it}^{0}\left(u_{it}^{z},u_{it}^{w}\right)=1-\Phi\left(X_{it}^{z}\gamma+\sum_{k=1}^{K}d_{ik}\sum_{j=1}^{J}\alpha_{jk}^{z}y_{ij}+u_{ijt}^{z}\right), \tag{8}$$

$$\begin{aligned} L_{it}^{1}\left(u_{it}^{z},u_{it}^{w}\right) &= \frac{1}{\sigma_w}\phi\left(\frac{w_{it}-X_{it}^{w}\delta-\sum_{k=1}^{K}d_{ik}\sum_{j=1}^{J}\alpha_{jk}^{w}y_{ij}-u_{it}^{w}}{\sigma_w}\right)\cdot \\ & \Phi\left(X_{it}^{z}\gamma+\sum_{k=1}^{K}d_{ik}\sum_{j=1}^{J}\alpha_{jk}^{z}y_{ij}+u_{ijt}^{z}\right), \end{aligned} \tag{9}$$

and $G(u_i \mid \Omega)$ is the joint normal density with covariance matrix Ω described in equation (12). While, in general, it is difficult to evaluate the multivariate integral in equation (5), it is straightforward to simulate the integral using well-known methods described in Stern (1997). The functional form of the conditional likelihood contribution associated with observed program choices, $L_i^y\left(u_i^y\right)$ in equation (6), follows from the assumption in equation (1) that the idiosyncratic errors are iid logit. The functional form of the conditional likelihood contribution for labor market outcomes, $L_{it}^{zw}\left(u_{it}^{z},u_{it}^{w}\right)$ in equations (7), (8), and (9), follow from the normality assumption for $\left(\upsilon_{it}^{z},\upsilon_{it}^{w}\right)$ and the bivariate normality assumption for $\left(\zeta_{it}^{z},\zeta_{it}^{w}\right)$ in equation (4). The log likelihood function is

$$L=\sum_{i=1}^{n}\log L_i.$$

In theory, the parameter estimates are consistent only as the number of independent draws used to simulate the likelihood contributions goes off to infinity. However, Börsch-Supan and Hajivassiliou (1992) shows that MSL estimates perform well for small and moderate numbers of draws as long as good simulation methods are used,[23] and Geweke (1992) shows that the simulation error occurring in simulation-based estimators is of order $(1/n)$ when antithetic acceleration is used.

4.2 Identification

There are two relevant notions of identification in this model. First, there is the general question of identification of model parameters in any nonlinear model. Second, service receipt and labor market outcome variables are likely to be endogenous. With respect to the first issue, covariation in the data between dependent variables and explanatory variables identifies many of the model parameters. For example, covariation between *male* and participation in training identifies the β_j coefficient in equation (1) associated with the *male* for j = *training*. Similarly, the covariation between *white* and employment

[23] We simulate all errors except for η and ε with antithetic acceleration (Geweke, 1992) and then compute likelihood contributions condition on the simulated errors. This is similar to simulation methods described in Stern (1992) and McFadden and Train (2000).

status identifies the γ coefficient in equation (2) associated with *white*, and the covariation between *white* and log quarterly earnings identifies the δ coefficient in equation (3) associated with *white*. Second moment parameters such as σ_ζ^2 and ρ_ζ in equation (4) are identified by corresponding second sample moments.

Two approaches are used to address the second identification problem. First, as in a difference-in-difference design, we control for pre-treatment labor market differences between those who do and do not receive services. If the differences in unobserved factors that confound inference in equations (2) and (3), u_{it}, are fixed over time, then controls for the observed pre-treatment labor market differences address the endogenous selection problem (see Meyer, 1995; Heckman, LaLonde, and Smith, 1999, Section 4). Second, we include two instruments in equation (1) that are excluded from equations (2) and (3). As described in Sections 3.3 and 8.2, our choice of instruments for service j is the propensity of an individual's counselor to assign other clients to service j and the propensity of an individual's field office to assign other clients to service j. Doyle (2007), Arrighi et al. (2010), Dean et al. (2010a, 2010b), and Clapp et al. (2010) use a similar instrument. In order for the instruments to be valid, it must be the case that a) they are correlated with service receipt, b) they do not belong directly in equations (2) and (3), and c) they are exogenous. While (a) is illustrated in Section 3.3 and (b) seems like a very reasonable assumption, (c) may be violated. For example, if there is significant unobserved variation in the availability of jobs where training is productive and counselors and field offices know that, such variation might affect average behavior of counselors and field offices and might also affect labor market outcomes. To some degree, we control for such variation by including measures of local labor market conditions directly in equations (2) and (3).[24] Also, there may be significant unobserved variation in the ability of counselors to match clients with jobs, thus affecting both his/her decisions about what type service to offer clients and later success in the labor market. We assume that these types of effects are not important in our analysis. Importantly, our approach for addressing the endogenous selection of services represents a substantial advance over the existing literature where the past research (often using RSA-911 data) generally relies on limited controls for pre-program earnings and assumes service participation is otherwise exogenous. Along with Aakvik, Heckman, and Vytlacil (2005) and Dean et al. (2010a, 2010b), this is the first study to identify the impact of VR services on labor market outcome using both a history of pre-program earnings and plausibly exogenous instrumental variables.

5 Estimation Results

5.1 Estimates

We divide up the discussion of parameter estimates into separate components. We begin by examining the estimated effect of services on labor market out-

[24] See Appendix 8.3 for details.

comes. Table 10 presents the estimates and associated standard errors for the effect of services on employment, and Table 11 presents the analogous results for log quarterly earnings. For each labor market outcome, the effects are allowed to vary across the six different service types and across different time periods relative to the initial service quarter. Given our rich labor market data, we are able to estimate both short-run (the first two years) and long-run (more than two years) effects of services and account for pre-service outcomes in the quarter prior to services as well as two or more quarters prior to the initial service. As noted in Section 4.2, inclusion of pre-treatment periods is a way to account for the effect of endogenous selection into services. This method of controlling for selection, which is the central idea of the difference-in-difference design, is used extensively in the literature (e.g., Meyer, 1995; Heckman et al., 1999). The quarter immediately prior to initial service provision is separated out because this quarter seems likely to have a distinct impact on selection and because of the well-documented variation in labor market behaviors just prior to the application period – the Ashenfelter dip (Ashenfelter, 1978; Heckman et al., 1999).

The first two columns of Tables 10 and 11, which display estimates for the quarters prior to the initial service, reveal evidence that selection is endogenous. Nearly all of the coefficients associated with periods two or more quarters prior to the initial service are substantial and statistically different than zero, the one exception being the coefficient on *education* in the *employment* equation. For *training*, the estimates reveal that those people with mental illness provided training services have lower pre-treatment employment probabilities but somewhat higher quarterly earnings. In other cases, such as *restoration*, the estimates imply selection is positively associated with pre-service labor market outcomes – people with mental illness with higher pre-treatment employment rates and earnings are more likely to be assigned to these services. In contrast, *maintenance* and *other services* are assigned to clients with relatively poor pre-service labor market outcomes. In general, the results for the quarter one period prior to services are qualitatively similar although in many cases are not statistically different than zero. Overall, these results suggest a complex and heterogeneous selection process where applicants are assigned to particular services based on underlying unobserved factors that are associated with pre-service labor market outcomes.

The last two columns of results display the estimated short- and long-run effects of services on labor market outcomes. These estimates should be interpreted relative to the coefficients associated with pre-service measures in the first two columns. For example, as seen in Table 10, prior to service provision, the effect of *training* on *employment* propensity is -0.264.[25] In the two years after the start of service provision, it rises to 0.295,[26] and then, in the longer

[25] Throughout this discussion, the effect of an explanatory variable on *employment* propensity means the partial derivative of the latent value associated with employment with respect to the explanatory variable.

[26] Recall that this 2-year period is one where those receiving services are in the program to various degrees and with varying durations.

Table 10: DRS Purchased Service Participation Effects on Employment Propensity

Variable	Two or More Quarters Prior to Service Participation	Quarter Prior to Service Participation	First 2 Years After Service Participation	More than 2 Years After Service Participation
Diagnosis & Evaluation	0.039 **	-0.216 **	-0.260 **	-0.466 **
	(0.010)	(0.076)	(0.016)	(0.009)
Training	-0.264 **	-0.100	0.295 **	0.180 **
	(0.013)	(0.095)	(0.020)	(0.010)
Education	0.022	-0.114	-0.291 **	-0.098 **
	(0.016)	(0.143)	(0.028)	(0.013)
Restoration	0.454 **	0.750 **	0.525 **	0.305 **
	(0.011)	(0.090)	(.020)	(0.010)
Maintenance	-0.532 **	-0.538 **	-0.398 **	-0.510 **
	(0.014)	(0.108)	(0.021)	(0.011)
Other Services	-0.140 **	-0.280 **	0.035	-0.062 **
	(0.013)	(0.119)	(0.023)	(0.011)

Notes:
1.Standard errors are in parentheses.
2.Single-starred items are statistically significant at the 10% level, and double-starred items are statistically significant at the 5% level.

run, it declines to 0.180. Relative to individuals in the sample who received no training services, the long-term *employment* propensity (z^*_{it} in equation (2)) is 0.180 higher for those that received training. The long-term effect of *training* on those who were trained after accounting for selection into service, is $0.180 + 0.264 = 0.444$.

The effects of each service type across the four time periods can be observed more easily in Figure 9. Relative to employment propensities two or more quarters prior to service provision, we observe that *training* and *other services* increase *employment* propensity while *diagnosis & evaluation* and *education* decrease *employment* propensity. *Restoration* and *maintenance* seem to increase *employment* propensity in the short run but either decrease or have no effect on *employment* propensity in the long run.

Table 11 displays estimates for the effect of purchased service provision on log quarterly earnings. As was true for *employment* propensity above, we allow for separate effects for each service type and for the same four periods. The same issues apply with respect to interpretation and identification except now the effects are on w_{it} in equation (3). The relative effects can be observed more easily in Figure 10. For earnings effects, with the exception of *diagnosis &*

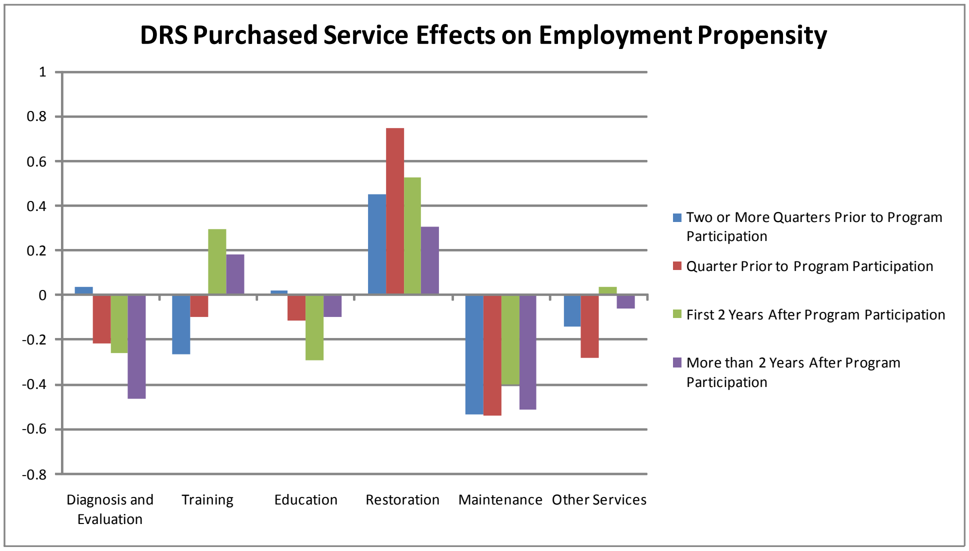

Figure 9: DRS Purchased Service Effects on Employment Propensity

evaluation, service receipt leads to higher earnings conditional on employment.[27] Dean and Dolan (1991) also find evidence of positive earnings effects in their earlier evaluation of VR services, although in some cases, especially for men, the results are not statistically significant. After using an instrumental variable to address the selection problem, Aakvik et al. (2005) find no evidence of employment effects of VR services in Norway.

Most previous evaluation of VR services focus on the impact of a single treatment indicator that is assumed to be conditionally exogenous. In this setting, the basic idea is to compare the differences in mean outcomes between treatment and control groups after conditioning on observed variables. For example, Figures 5 and 6 above, which display the unconditional mean employment and earnings outcomes respectively, reveal little pre-program differences, fairly substantial positive post-treatment employment associations, and almost no relationship between treatment and earnings. The structural model estimated in this paper extends this approach in several important ways: first, by conditioning on observed covariates; second, by accounting for six different types of service rather than a single treatment indicator; and finally, by using instrumental variables in a model with endogenous service provisions. The results from the structural model estimates presented in this section suggest a much more complex and nuanced story, with evidence of pre- and post-program labor market differences that vary across services, estimated employment effects that are positive for some services and negative for others, and estimated earnings

[27] With the exception of *diagnosis & evaluation*, all F-statistics testing for the joint significance of the short-term and long-term *log quarterly earnings* effects relative to the effect prior to program participation are statistically significant with p-values less than 0.0001.

Table 11: DRS Purchased Service Participation Effects on Log Quarterly Earnings

Variable	Two or More Quarters Prior to Service Participation	Quarter Prior to Service Participation	First 2 Years After Service Participation	More than 2 Years After Service Participation
Diagnosis & Evaluation	-0.189 **	-0.512 **	-0.302 **	-0.209 **
	(0.014)	(0.096)	(0.024)	(0.012)
Training	0.035 **	0.059	0.032	0.108 **
	(0.017)	(0.137)	(0.029)	(0.013)
Education	0.122 **	-0.016	-0.049	0.248 **
	(0.021)	(0.156)	(0.039)	(0.016)
Restoration	0.148 **	0.057	0.279 **	0.325 **
	(0.016)	(0.111)	(0.028)	(0.014)
Maintenance	-0.570 **	-0.253 *	-0.457 **	-0.372 **
	(0.017)	(0.139)	(0.032)	(0.014)
Other Services	-0.131 **	-0.313	0.024	0.086 **
	(0.017)	(0.149)	(0.031)	(0.014)

Notes:

1.Estimates are effects on log quarterly earnings conditional on employment.

2.Standard errors are in parentheses.

3.Single-starred items are statistically significant at the 10% level, and double-starred items are statistically significant at the 5% level.

effects that, except for *diagnosis & evaluation*, are consistently positive.

Figure 11 uses the *employment* effects from Table 10 and the *log quarterly earnings* effects from Table 11 to compute the average marginal effect of each service type on labor market outcomes. In particular, for each service j with value defined in equation (1) and each labor market outcome z_{it} and w_{it} defined in equations (2) and (3) respectively, we compute

$$\frac{1}{n}\sum_{i}\left[v_{ijk}\left(1\right)-v_{ijk}\left(0\right)\right]$$

where

$$v_{ijk}\left(y_{ij}\right)=\left[\frac{1}{T_{ki}}\sum_{t\in\Im_{ki}}v_{ijkt}\left(y_{ij}\right)\right]-\left[\frac{1}{T_{0i}}\sum_{t\in\Im_{0i}}v_{ijkt}\left(y_{ij}\right)\right];$$

$v_{ijkt}\left(y_{ij}\right)$ is the outcome measure, employment (measured in probability increments) or quarterly earnings conditional on employment (measured in \$1000), for person i at time t conditional on whether service j is received y_{ij} (with no other service being received); $\Im_{ki}$ is the set of quarters observed in the data for observation i before service receipt excluding the quarter preceding service ($k=0$), in the short run ($k=1$) or the long run ($k=2$); and T_{ki} is the number of quarters in $\Im_{ki}$. For example, for *training*, in Table 10 and Figure 9,

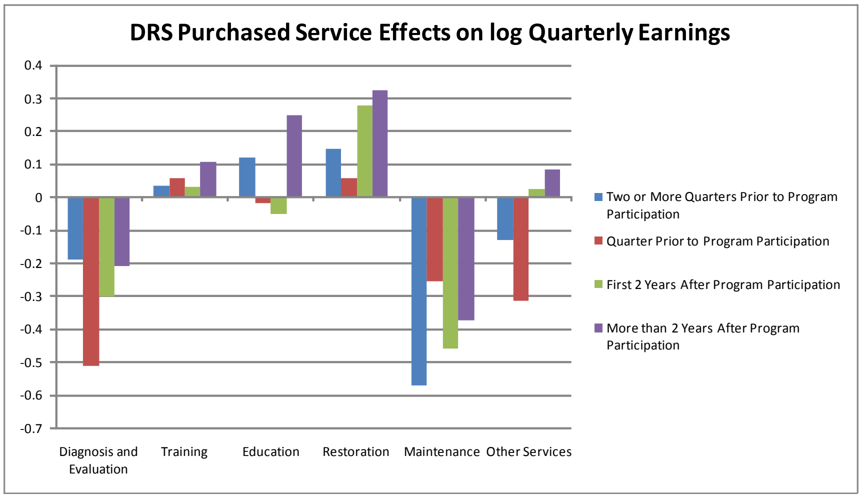

Figure 10: DRS Purchased Service Effects on log Quarterly Earnings

we see that, relative to the quarters preceding service, *training* increases *employment* propensity (equation (2)) by $0.295 + 0.264 = 0.559$ in the short run, causing average short-run employment probabilities to increase by 12.9%;[28] a similar calculation implies long-run employment probabilities increase by 10.2%. In Figure 10, we see that, relative to the quarters preceding service, *training* slightly decreases quarterly earnings in the short run and increases them in the long run.[29] These are the four outcomes associated with *training* presented in Figure 11.[30] One can see that, with the exception of *diagnosis & evaluation* and *education*, all services generally have positive labor market outcome effects.[31]

The effects hardest to understand are for *diagnosis & evaluation* with negative outcomes in the short- and long-run for employment and mixed results for conditional earnings. There is no reason to think that the receipt of purchased diagnosis & evaluation services would lead directly to poor labor market outcomes. However, purchased diagnosis & evaluation services differ from other types of service in a number of ways, some of which imply that receipt of such

[28] Nonlinearity in the transformation of expected latent variables into conditional probabilities causes the difference between 0.559 and 0.129.

[29] The transformation of log wage effects into wage effects causes deviations in the results between Table 11 and Figure 10.

[30] A complete set of moments and extrema associated with all parameter estimates is available at Stern (2010).

[31] The short-run effects may be particularly sensitive to the way we classify service receipt. Recall that we measure labor market outcomes relative to the first quarter in which services are provided.

One might think that education services are typically provided over many quarters causing our short-run comparisions to reflect on-going service receipt rather than negative employment effects. However, the distribution of the length of education service receipt is about the same as the other services.

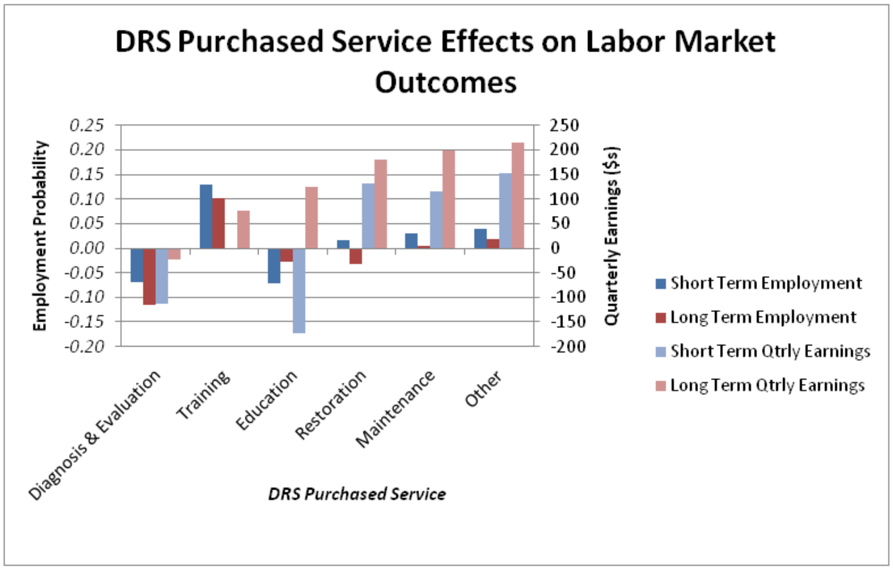

Figure 11: DRS Purchased Service Effects on Labor Market Outcomes

services acts very much like a selection effect:

1. One purpose for diagnosis & evaluation services is to determine program eligibility. As such, every applicant receives them. However, over half the sample received these services in-house rather than by purchase (see Table 2), and most in-house services are for diagnosis & evaluation services. Why that might be is something we do not observe directly in the data.

2. It may be that purchased diagnosis & evaluation services are for clients with especially difficult cases implying that receipt of such services is very much like a selection effect. It might be noted in this regard that 79% of those receiving diagnosis & evaluation services by purchase have exams which may involve a specialist. However, this is equivalent to stating that *diagnosis & evaluation* is endogenous; it is not clear, why our counselor and office instruments are not controlling for such endogeneity.

3. It may also be that clients who receive purchased diagnosis & evaluation services are more likely to be diagnosed with problems that make it difficult for them to succeed in the labor market, and the DRS counselor influences them to move in a different, more rewarding direction. In such cases, while this would not look like a success in our data, in reality, it might lead to the most productive outcome available.

4. It might be the case that the least successful counselors with respect to diagnoses & evaluation are both the most likely to use purchased diagnosis

& evaluation services and are the least likely to succeed in helping their clients.

Differentiating among these and possibly other explanations is left to future work. None of these issues apply to the other five aggregated services.

Because of the variation in effects over time and over labor market outcomes seen in Figure 11, it is difficult to infer the long-run benefits of each service. Figure 12 reports the mean present value for 10 years of earnings flows (measured in \$1000) excluding service costs , a 95% confidence range,[32] and the minimum and maximum present value of each service.[33] Except for *diagnosis & evaluation*, all of the services have positive long-run benefits. On average, *training*, *restoration*, and *other services* have benefits on the order of \$7700, \$8600, and \$8700 respectively, while *education* and *maintenance* have positive benefits of \$1300 and \$2600 respectively. It should be noted that, in Figure 11, *education* has a negative effect on both short- and long-run employment probabilities but a substantial long-run positive effect on quarterly earnings conditional on employment. Figure 12 shows that the long-run conditional earnings effects essentially offset the negative employment effects for present value calculations. One other notable feature of the discounted benefits calculations illustrated in Figure 12 is the high degree of variability across the caseload. The discounted benefits associated with training services, for example, range from \$500 to nearly \$35000. For the other service categories, there are notable fractions of the caseload that would receive negative benefits.

Table 12 provides estimates of the effects of various demographic characteristics on the propensity to use different services (y^*_{ijt} in equation (1)). For the most part, these observed characteristics do not have statistically significant effects on service receipt. Six characteristics have statistically significant impacts on the receipt of a single service – e.g., *education* positively impacts the receipt of educational services (0.082) – and another six attributes affect the probability of receiving two services, including the negative impact of having a mental illness on receiving *diagnosis & evaluation* (−0.618) and *training* (−0.950). This last finding might reflect the fact that many referrals of persons with mental illness to VR agencies come from sources such as community service boards, where extensive diagnosis & evaluation has already been conducted. Interestingly, however, there is no statistically significant effect associated with having a serious mental illness or a significant disability. Finally, three variables – *# dependents*, *learning disability*, and *has driver's license* – have statistically significant impacts on receiving three service types. In the latter case, the ability to drive decreases the likelihood of receiving maintenance (e.g., a bus ride to an appointment) but improves the likelihood of receiving diagnosis & evaluation and education services.[34]

Table 13 presents estimates of counselor and office effects as defined in Ap-

[32] The 95% confidence range provides information about the variation in benefits across individuals caused by the nonlinearity of the model and variation in other explanatory variables.

[33] We use a quarterly discount factor of 0.95. Because the distribution of benefits is highly skewed, the normal approximation is not appropriate. Instead, we report the 0.025 and 0.975

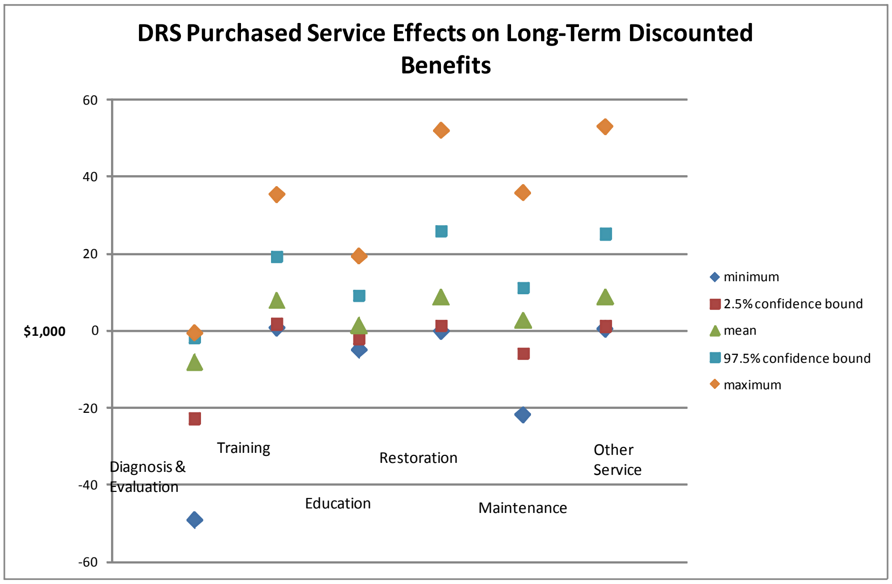

Figure 12: DRS Purchased Service Effects on Long-Term Discounted Benefits

pendix 8.2. There are two types of coefficient estimates reported in the table: a) the counselor and office effects and b) the missing counselor effects. The counselor and office effects should be interpreted as $\partial E y^*_{ij}/\partial e_i$ where y^*_{ij} is the latent variable associated with receipt of service j in equation (1) and e_i is the counselor or office effect defined in Appendix 8.2; note that these are restricted to be the same across different services. The missing counselor effects are the effect on y^*_{ij} when the relevant counselor does not have enough other clients to compute a set of counselor effects.[35] These counselor and office instrumental variables turn out to have large and statistically significant effects on service provision across clients. One should note that, in Table 12, we are controlling for a pretty full set of demographic characteristics. So it is unlikely that these results reflect variation in the mix of clients across counselors and/or field offices.[36] On the one hand, arbitrary variation in service provision across counselors and field offices is good in that it allows for DRS to evaluate different service programs more effectively by creating exogenous variation in service provision. On the

quantiles of the empirical distribution.

[34] The *education missing* variable is statistically significantly negative across all services as well. It turns out that almost all of the individuals with *education missing* were closed during the application process. Thus, in an important sense, causation for this variable runs the other way.

[35] We allow missing counselor effects to vary over services. However, we restrict missing office effects coefficients to be zero because there are not enough cases and those that exist are too highly correlated with missing counselor effects to estimate both with any precision.

[36] When assigning clients to counselors, consideration is given to client disability, counselor load, etc. Clients have no input in their assignment to a counselor.

Table 12: Effects of Client Characteristics on Service Receipt by Type

Variable	Diagnosis & Evaluation	Training	Education
Constant	0.696	1.234 *	-1.218
Male	-0.065	-0.317 *	-0.088
White	-0.016	-0.053	-0.198
Education	-0.018	-0.005	0.082 **
Special Education	0.436	-0.529	0.107
Education Missing	-1.051 **	-3.917 **	-2.670 **
Age/100	0.146	-0.234	-0.110
Married	-0.198	-0.314	-0.060
# Dependents	-0.040	-0.214 **	-0.095
Transportation Available	0.173	0.146	0.503 *
Has Driving License	0.251 *	-0.266	0.690 **
Receives Govt Assistance	0.486 **	0.758 **	0.247
Musculoskeletal Disability	0.259	-0.478 **	0.072
Learning Disability	0.679 **	0.049	-0.657
Mental Illness	-0.618 **	-0.950 **	-0.402
Substance Abuse	0.033	-0.389 *	0.281
Disability Significant	0.316	0.118	-0.386
Disability Most Significant	0.440 *	0.480	-0.468
SMI	-0.029	0.653	0.422
Male * SMI	-0.131	0.142	-0.566
White * SMI	-0.126	0.185	0.444
Education * SMI	0.007	-0.064	-0.002
Age/100 * SMI	-0.021	0.309	-0.424

Table 12 (continued): Effects of Client Characteristics on Service Receipt by Type

Variable	Restoration	Maintenance	Other Services
Constant	0.658	1.666 **	0.008
Male	-0.282 *	0.041	-0.092
White	-0.120	-0.326 *	-0.217
Education	-0.002	-0.029	0.011
Special Education	0.095	-0.847 *	-0.096
Education Missing	-1.304 **	-3.787 **	-2.437 **
Age/100	-0.009	-0.389 *	0.072
Married	-0.304 *	-0.447 *	-0.295
# Dependents	0.123 *	-0.014	-0.130 *
Transportation Available	-0.060	-0.214	0.370 *
Has Driving License	0.114	-0.319 *	0.085
Receives Govt Assistance	-0.164	0.363	0.073
Musculo/Skeletal Disability	-0.143	0.130	-0.012
Learning Disability	2.134 **	-0.054	0.545 *
Mental Illness	-0.370	-0.503	-0.547
Substance Abuse	-0.098	0.340 *	-0.207
Disability Significant	0.093	-0.007	-0.040
Disability Most Significant	0.342	0.355	0.120
SMI	-1.447	-0.596	0.526
Male * SMI	0.269	-0.226	0.111
White * SMI	-0.003	0.144	0.029
Education * SMI	0.033	-0.052	-0.032
Age/100 * SMI	-0.188	0.943 **	-0.059

Notes:
1.Standard errors not presented to save space but are available from the corresponding author.
2.Single-starred items are statistically significant at the 10% level, and double-starred items are statistically significant at the 5% level.

Table 13: Counselor and Office Effects on Service Receipt

	Estimate		Std Err
Counselor Effect	0.330	**	0.103
Office Effect	0.777	**	0.068
Missing Counselor Effects			
Diagnosis & Evaluation	-0.475		0.347
Training	0.123		0.404
Education	-0.732	*	0.433
Restoration	-0.555	*	0.355
Maintenance	-0.308		0.414
Other Services	-0.349		0.406

Notes:
1.Single-starred items are statistically significant at the 10% level, and double-starred items are statistically significant at the 5% level.
2.Other than those reported, missing counselor and field office effects parameters were excluded because of multicollinearity problems.

other hand, it is not clear that it is in the best interest of the present clients to have field offices and counselors have such strong influence on the provision of services.

Table 14 reports the effects of the demographic, socioeconomic, and disability-related characteristics on the two labor market outcomes of interest (z^*_{it} in equation (2) and w_{it} in equation (3)). Almost all of the estimates are statistically significant. Many of the estimates are as expected including positive effects of being white on *employment* propensity (0.227) and *log quarterly earnings* (0.417) as well as positive effects of *education* on *employment* propensity (0.034) and *log quarterly earnings* (0.056). The two transportation variables also have positive impacts on both labor market outcomes. The two local labor market conditions do not perform very well; none are significant, and they have opposite signs. This probably is caused by the high correlation of the two measures. Some of the demographic and socioeconomic parameter estimates are counter-intuitive. In particular, being male decreases *employment* propensity (-0.066), receipt of special education increases both *employment* propensity (0.150) and *log quarterly earnings* (0.415), while being married decreases both *employment* propensity (-0.301) and *log quarterly earnings* (-0.219). The positive effects for *special education* may be due to the higher propensity for such people to use long-term employment support services (LTESS).[37] We are in the process of

[37] Sheltered employment involves working, potentially at sub-minimum wages, in a segregated employment setting with other persons with disabilities. Supported employment involves competitive employment in an integrated work setting, usually with the assistance of

collecting LTESS data to be used in subsequent research.

The diagnosis of a mental illness in the "base case" versus being initially diagnosed with mental illness in a subsequent application for VR services has a negative effect on *employment* propensity (-0.115) while increasing *log quarterly earnings* (0.437). Meanwhile, the disability severity-related variables have the expected signs, with negative effects of *significant* and *most significant* disabilities (relative to mild) on both labor market outcomes. Unlike its impact on service provision as seen in Table 12, the *SMI* estimates are explaining a significant amount of variation in labor market outcomes. *SMI*, by itself, reduces employment ($-.275$) but increases *log quarterly earnings* (0.390). For males and whites, there are added interaction effects, all adversely affecting labor market outcomes. *Education* and *age* interacted with *SMI* also have small but statistically significant effects on outcomes. Baldwin (2005) estimates the effect of mood disorder, anxiety disorder, and adjustment disorder on employment probabilities and finds an average reduction in employment probability on the order of 0.3 (see McKeithen and Stern, 2007, for calculations). Our estimates imply smaller effects, at least for significant mental health problems similar to those considered by Baldwin. A big part of the reason for this is probably that our sample consists only of people who have been identified as having a mental health problem while Baldwin (2005) uses the SIPP sample.

Our model has a rich error covariance structure, as seen in equation (4). This allows for the possibility that unobservables associated with service provision are correlated with unobservables associated with labor market outcomes. The factor loadings for *Factor 1* in Table 15 demonstrate positive correlations between the errors associated with the provision of almost all service types and the error associated with *employment* propensity. However, the correlation between the errors for *employment* propensity and *log quarterly earnings* is negative (0.898 and -0.094). This suggests that there is some unobserved personal characteristic, maybe desire to work, that increases service provision probabilities and employment probabilities but decreases reservation wages, thus decreasing earnings.

By contrast, the factor loadings for *Factor 2* imply small, statistically insignificant correlations between errors associated with service provision and the errors associated with labor market outcomes but positive and significant correlations between the errors associated with *employment* propensity (0.426) and *log quarterly earnings* (0.255). This suggests another unobserved characteristic, perhaps ability, increasing *employment* propensity and *log quarterly earnings* but having no real impact on service receipt.

The estimates of the other elements of the error structure are reported in Table 16. The serial correlation estimate ρ_η is very large due to the high degree of inertia associated with labor market spells seen in Figure 8. The correlation between the two different labor market outcome errors ρ_ζ is also large; the estimate of ρ_ζ and the *Factor 2* factor loadings for the two labor

a job coach. See Kregel and Dean (2002) for more details. Of those not receiving *special education*, 9.8% receive LTESS, and, of those receiving *special education*, 28.6% receive LTESS; It is clear that *special education* is an important predictor of LTESS receipt.

Table 14: Labor Market Effects

Variable	Employment Estimate		Employment Std Err	Log Quarterly Earnings Estimate		Log Quarterly Earnings Std Err
Constant	0.098	**	0.031	5.244	**	0.035
Male	-0.066	**	0.008	0.361	**	0.009
White	0.227	**	0.009	0.417	**	0.010
Education	0.034	**	0.001	0.056	**	0.001
Special Education	0.150	**	0.030	0.415	**	0.036
Education Missing	0.273	**	0.017	0.505	**	0.023
Age/100	-0.375	**	0.009	0.240	**	0.011
Married	-0.301	**	0.010	-0.219	**	0.010
# Dependents	0.047	**	0.003	0.096	**	0.003
Transportation Available	0.088	**	0.009	0.089	**	0.011
Has Driving License	0.333	**	0.010	0.413	**	0.011
Receives Govt Assistance	-0.581	**	0.012	-0.451	**	0.012
Musculoskeletal Disability	0.057	**	0.009	0.080	**	0.011
Learning Disability	0.242	**	0.016	0.347	**	0.017
Mental Illness	-0.115	**	0.017	0.437	**	0.018
Substance Abuse	0.335	**	0.010	0.213	**	0.011
Disability Significant	-0.319	**	0.012	-0.417	**	0.013
Disability Most Significant	-0.409	**	0.013	-0.530	**	0.014
SMI	-0.275	**	0.042	0.391	**	0.039
Male * SMI	0.231	**	0.019	-0.370	**	0.019
White * SMI	-0.492	**	0.019	-0.358	**	0.019
Education * SMI	-0.023	**	0.002	-0.043	**	0.002
Age/100 * SMI	0.443	**	0.022	0.031		0.021
Local Employment Rate	0.158		0.231	0.228		0.273
Metro Employment Rate	-0.268		0.232	-0.258		0.273

Note: Single-starred items are statistically significant at the 10% level, and double-starred items are statistically significant at the 5% level.

Table 15: Covariance Factor Loadings

Variable	Factor 1 Estimate		Factor 1 Std Err	Factor 2 Estimate		Factor 2 Std Err
Diagnosis & Evaluation	0.159	**	0.067	0.041		0.067
Training	0.196	**	0.080	-0.096		0.082
Education	0.292	**	0.110	-0.236	**	0.115
Restoration	-0.013		0.078	0.019		0.079
Maintenance	0.278	**	0.082	0.065		0.088
Other Services	0.260	**	0.089	0.244		
Employment	0.898	**	0.005	0.426	**	0.005
Log Quarterly Earnings	-0.094	**	0.003	0.255	**	0.004

Notes:

1. Single-starred items are statistically significant at the 10% level, and double-starred items are statistically significant at the 5% level.
2. The identifying condition associated with the factor loadings is that the factor loadings for the six different services are orthogonal. We impose this condition by computing the factor loading for factor 2 on other services as a function of the other 11 relevant factor loadings. The factor loadings associated with labor market outcomes are not part of the orthogonality condition.

Table 16: Other Covariance Terms

Variable	Estimate	Std Err	Variable	Estimate	Std Err
ρ_η	0.981 **	0.000	ρ_ζ	0.825 **	0.004
σ_ζ	0.186 **	0.001	σ_w	1.212 **	0.002

Notes:

1.Double-starred items are statistically significant at the 5% level.

2.Correlation terms are estimated using the transformation,

ρ=[2exp{ς}/[1+exp{ς}]]-1

where ς is estimated to insure that -1<ρ<1. Standard deviations are estimated using the transformation, σ=exp{ς} where ς is estimated to insure that σ>0. Standard errors for both are derived using the delta method.

market outcomes in Table 15 imply a high degree of correlation between the two errors. The estimate of the log earnings error σ_w is quite large, implying that a standard deviation in quarterly earnings due to unobserved factors is on the order of \$7024.[38] It is unclear how much of this variation is due to variation in wages and how much is due to variation in hours. Baldwin (2005) finds wage effects on the order of -0.2 (see McKeithen and Stern, 2007, for calculations) but does not estimate hours effects.

5.2 Specification Tests

We use standard goodness-of-fit tests to measure how well we are predicting service provision probabilities. For each service, we decompose the sample into 40 cells, each of length 0.025, stratified by the predicted probability of service receipt.[39] Then we construct the standard χ^2 test statistic. The results are reported in Table 17. For each service, we do a good job of predicting service receipt probabilities as evidenced by all of the negative normalized statistics. For service provision probabilities, we accept the null that the model predictions equal observed probabilities at the 5% percent significant level.

We perform the same test for employment probabilities disaggregated into probabilities before and after service receipt.[40] The test statistics are $\chi^2_{33} =$

[38] The variance of a lognormal random variable with parameters (μ, σ^2) is

$$\exp\left\{2\left(\mu+\sigma^2\right)\right\} - \exp\left\{2\mu+\sigma^2\right\}.$$

Setting $\mu = 7.52$ (from Table 7) provides the result.

[39] For each test, some cells are empty and therefore not used.

[40] *Before service receipt* includes the quarter before receipt, and *after service receipt* includes both quarters in the first two years after receipt and the longer run.

Table 17: Overall χ^2 Goodness-of-Fit Statistics for Service Probabilities

Service	χ^2	DF	Normalized Statistic
Diagnosis & Evaluation	9.71	30.00	-2.62
Training	22.61	35.00	-1.48
Education	27.51	23.00	0.66
Restoration	29.90	37.00	-0.83
Maintenance	17.66	34.00	-1.98
Other	26.79	28.00	-0.16

Note: Normalized statistic is $(\chi_k^2\text{-}k)/(2k)^{1/2}\sim N(0,1)$ where k is the degrees of freedom (DF).

203.7 for employment probabilities before service receipt and $\chi^2_{32} = 1405.7$ for employment probabilities after service receipt. Both of these are highly significant implying a poor fit. Figure 13 plots the deviations between predicted and sample employment probabilities for the two periods. Deviations between the 45° line and the other two *sample lines* at any particular predicted probability represent that part of employment probability that we are not predicting. The model does a pretty good job predicting employment probabilities in the *before* period up to a predicted probability of about 0.6, after which the fit worsens. For example, the predicted probability of 0.8 exceeds the sample probability by close to 0.3. For the *after* period, the model does a pretty good job until a predicted probability of about 0.25 at which point the model overpredicts employment rates up until a predicted probability of about 0.82. In some regions, the fits can be off by as much as 0.3. However, overall, we are basically grouping individuals into the correct range of employment probabilities.

We consider a number of Lagrange Multiplier (LM) tests to test for missing pieces of our model – the interaction of demographic characteristics, the interaction of service type in the employment outcome equations, and the duration of service provision by type. In each case, we report two types of statistics. Let $\log L_i(\theta, \xi)$ be the log likelihood contribution from equation (5) where θ is the vector of parameters defined above equation (5) and ξ is the vector of parameters associated with a particular hypothesis test; i.e., $H_0 : \xi = 0$ vs $H_A : \xi \neq 0$ which, in the context of the LM test, becomes $H_0 : \partial \log L_i(\theta, \xi)/\partial\xi = 0$ vs $H_A : \partial \log L_i(\theta, \xi)/\partial\xi \neq 0$. The standard LM test statistic is

$$\left[\sum_i \frac{\partial \log L_i(\theta,\xi)}{\partial \xi}\right]' D^{-1} \left[\sum_i \frac{\partial \log L_i(\theta,\xi)}{\partial \xi}\right] \left[\sum_i \frac{\partial \log L_i(\theta,\xi)}{\partial \xi}\right] \sim \chi^2_k \quad (10)$$

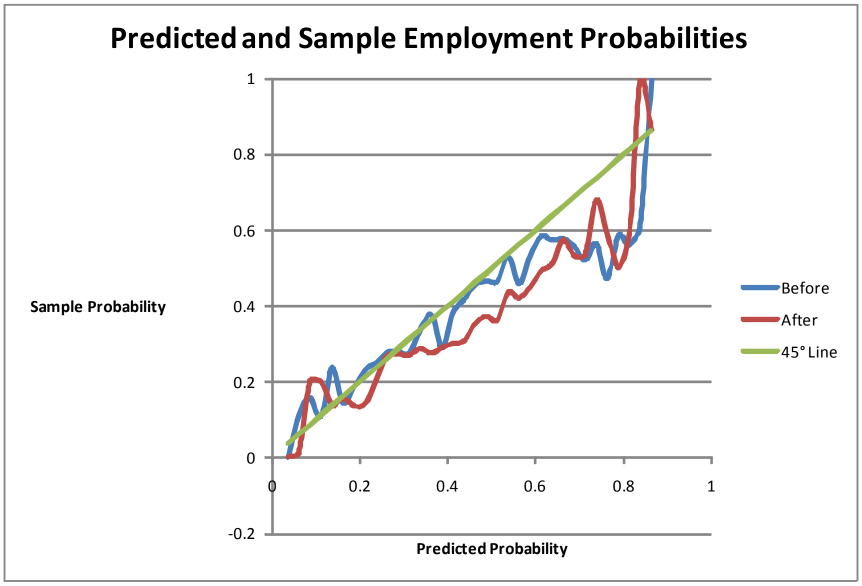

Figure 13: Predicted and Sample Employment Probabilities

where $D^{-1}[\cdot]$ is the inverse covariance matrix of its argument and k is the number of elements in ξ. However, for many of the hypotheses considered, $D^{-1}[\cdot]$ is not well-behaved because the score statistics are too colinear. Thus, we also report t-statistics,

$$\left[\sum_i \frac{\partial \log L_i(\theta,\xi)}{\partial \xi}\right] / \sqrt{Var\left[\sum_i \frac{\partial \log L_i(\theta,\xi)}{\partial \xi}\right]} \sim N(0,1) \tag{11}$$

for each individual element of ξ.[41]

In Tables 12 and 14, we include *marry* and *# dependents* as explanatory variables but do not allow for interactions between them and *male*. At least for labor market outcomes, much of the literature (e.g., Ettner, Frank, and Kessler, 1997; Kimmel and Kniesner, 1998) suggests such an interaction. Unfortunately, for service choices, there is not enough variation in the data with respect to *marry* and *# dependents* conditional on *male* $= 1$ to perform a valid overall test. To some degree, this occurs for our population because the marriage rate is significantly lower than in the general population of adults. For labor market outcomes, the interaction terms have statistically insignificant negative effects for both *marry* and *# dependents* with $\chi^2_4 = 5.76$.

Next, we consider allowing for interactions among pairs of services in the labor market outcome equations. It should be noted that the nonlinearity of the

[41] It should be noted that $D^{-1}[\cdot]$ in equation (10) depends on the covariance matrix of all of the parameter estimates (θ,ξ), while the denominator in equation (11) is independent of covariances with other elements of (θ,ξ).

model implies a certain amount of interaction. However, it may not be appropriate to rely strictly on the model structure to deal with much more complex interactions. In fact, none of the individual t-tests are statistically significant, but $\chi^2_{30} = 72.52$ is statistically significant at the 5% level. This suggests that a more parsimonious specification for interactions might be appropriate. We consider two such specifications. First, for each labor market outcome, we consider the addition of a dummy variable that is equal to one iff the individual uses at least two services. Neither of the labor market effects are statistically significant, and $\chi^2_2 = 4.02$ is not statistically significant. Second, we distinguish between *diagnosis* & *evaluation* versus the other five services. In particular, for each labor market outcome, we include both the dummy in the first case, and we add a second dummy equal to one iff the individual uses at least two services excluding *diagnosis* & *evaluation*. Again, none of the results are statistically significant. Overall, these results suggest that it is not important to allow for service interactions.

Next, based on comments we received from counselors and program administrators, we test to see if the effects of services vary across gender or race. We included interactions for *male* and *white* and then with each of the 6 service types, all allowed to vary across effects before service, short-run after service, and long-run after service. The results are reported in Table 18. Most of the mean score statistics are statistically insignificant (and therefore not reported in Table 18). Also, no score statistics for *diagnosis* & *evaluation*, *training*, or *maintenance* are statistically significant. For *education*, relative to black women (the base group), white women have larger short-run and long-run earnings effects. For *other services*, relative to black women, black men have larger short-run and long-run employments effects, and white men have a smaller long-run employment effect.[42] Overall, the χ^2_{108} test statistic is very large, but this is almost totally caused by correlations of score statistics between these interaction terms and the other estimated model parameters; it does not appear that allowing for such interactions is qualitatively important.

Finally, in a first attempt to measure the potential effect of a continuous service variable, we test to see if the length of service[43] affects labor market outcomes. None of the t-statistics are statistically significant, but, for the joint test, $\chi^2_{12} = 21.05$ is statistically significant at the 5% level. There are two possibly counteracting effects here: true positive value for longer service receipt and selection effects due to clients with greater needs receiving service longer and then not performing as well in the labor market. The same may (or may not) be true for service expenditures. We leave this question for future work.

6 Return on Investment

The preceding analysis suggests that, except for purchased diagnosis & evaluation services, observed DRS services have long-run positive effects on labor

[42] Also, for *restoration*, relative to black women, white women have larger selection effects.
[43] Length of service is defined in footnote 13 as Δ_e.

Table 18: Statistically Significant Score Statistics Associated with Lagrange Multiplier Test for Gender/ Race/ Service Interactions

Market Outcome	Male	White	Service	Mean Score
Short-run, Employment	No	Yes		0.141
Long-run, Employment	Yes	No		-0.471
Long-run, Employment	Yes	Yes		-1.022
Before, Employment	No	Yes	Restoration	0.078
Short-run, Employment	Yes	No	Other Service	0.271
Long-run, Employment	Yes	No	Other Service	0.302
Long-run, Employment	Yes	Yes	Other Service	-0.218
Short-run, Earnings	No	Yes	Education	0.022
Long-run, Earnings	No	Yes	Education	0.013
Short-run, Earnings	No	Yes	Other Service	-0.063

Notes:

1.All reported means are statistically significant at the 5% level.

2.The overall χ_{108}^2 test statistic is 64680.4 which is statistically significant at any conventional level.

market outcomes (see Figures 11 and 12). In this section, we examine the social welfare implications of VR services by comparing the estimated benefits and costs of the program. The primary monetary benefits and costs of VR services are estimated using our model and the DRS data on the costs of purchased services. There are, however, many factors for which we do not have direct evidence on the associated costs and benefits. For example, the costs of services provided a) internally or b) as similar benefits are not observed in the DRS data file. For these items, we present more speculative evidence on the qualitative and, when possible, quantitative impact.

We simulate the private labor market benefits to DRS clients using the structural model estimates summarized in Section 5.[44] In particular, we compute the mean present discounted value of the provided services relative to receiving no services using both a 5- and 10-year post-treatment observation period for those individuals who received some service. The estimated mean discounted benefits are \$2614 with a standard deviation of \$5619 using the 5-year window and \$5728 with a standard deviation of \$10975 using a 10-year window. Thus, the mean long-run discounted benefits are more than twice the mean short-run

[44]This simulation has a similar structure to the one used to compute marginal effects in Section 5.1 (see Figure 12). But here we compute the present discounted value of the actual treatments provided by DRS rather than a conjectured treatment for single service, j. Formally, we first compute the short- and long-run effect of the program for each individual:

$$\Delta_i = v_{ik}(y_i) - v_{ik}(0)$$

where $v_{ik}(y_i)$ is the estimated labor market earnings under the realized services y_i and $v_{ik}(0)$ is the estimated earnings that would be observed if no services were provided.

benefits, reflecting the effect of adding an extra 5 years and the fact that the long-run (over 2 years) labor market effects are estimated to be much larger than the short-run effects and are assumed for the purpose of this analysis to last throughout the 10-year window. If we exclude *diagnosis* & *evaluation* from the analysis (for reasons already discussed), then the estimated mean discounted benefits are $7256 with a standard deviation of $6670 using the 5-year window and $14283 with a standard deviation of $13415 using a 10-year window.

While these estimated benefits are derived directly from the structural model, there are several reasons they may not reflect the true social benefits of VR services. First, some of the estimated earning benefits may reflect the displacement of non-VR participants, particularly if VR services do not improve the VR participant skills or the job matching process. In general, however, training programs for low-skilled workers are not thought to cause notable labor market displacements (see Lalonde, 1995). Second, VR services may lead to other social benefits associated with the increased attachment to the labor market and the resulting reduction in use of the social welfare system. While society does not benefit from reduced transfer payments or increased tax revenues – taxpayer gains exactly offset VR participant losses (except for changes in deadweight loss) – social benefits may result from reduced administrative cost associated with welfare programs and increased VR participant utility due to reduced welfare dependence (Lalonde, 1995). Finally, there is substantial heterogeneity in the discounted benefits across the VR participants, suggesting that there may be a great deal of variation in the overall benefits estimates (see Figure 12). Many of the clients are estimated to have negative benefits from VR services.

As noted in Section 3, DRS services are provided in any combination of three ways: a) internally by DRS personnel, b) as a similar benefit (i.e., purchased or provided by another governmental agency or not-for-profit organization with no charge to DRS), and/or c) as a purchased service through an outside vendor using DRS funds. The DRS data report purchased services but not in-house services or similar benefits. Table 19 displays the mean costs of purchased services for each service. On average, mean expenditures for training, at $600, account for just over 40% of the total average cost of purchased services. Interestingly, the average cost for training is substantially less than the mean long-term discounted marginal benefit of $7700 (see Figure 12). Overall, the mean costs of purchased services among all 1555 clients (30% of whom receive no purchased services) equals $1436 with a standard deviation of $3353. These mean cost estimates have not been discounted, and thus will be inflated to the extent the purchased services are provided over long periods.

The DRS data do not provide information on the costs of DRS-provided services, similar benefits, or of administrating the program. To estimate these costs, we use information on DRS spending by fiscal year as reported to the US Social Security Administration. These reports summarize information on aggregate administrative costs, DRS-provided counseling, guidance, and placement service costs, purchased service cost, and size of the caseload for each fiscal year. These reports, however, do not provide information on the costs associated with similar benefits. While there is some variation in the distribution of costs across

Table 19: Moments of Expenditure Data on Purchased Services

Service	Mean Expenditure	Std Deviation	% with Positive Expenditure
Diagnosis & Evaluation	$126	$415	39%
Training	$601	$1,956	29%
Education	$193	$1,096	11%
Restoration	$227	$771	28%
Maintenance	$246	$1,320	25%
Other Services	$44	$551	6%

Notes:
1.Moments are not conditional on receipt of service.
2.Moments do not include cost of in-house services or similar benefits.

years, in general, non-purchased service and administrative costs account for 45% of total expenditures, reflecting an average cost per client of roughly \$200 per month.

While these reports do not provide information specific to the different impairment groups, this auxiliary information can be used to infer the cost for our sample of applicants with mental illnesses. Two different approaches are used. In the first, we anchor on the fact that purchased services account for 45% of total VR costs. Given that purchased service costs for our sample average \$1436 per client, fixed costs are estimated to be \$1800 ($\approx$ (\$1436/0.45) $-$ \$1436) per client. In the second, we anchor on the fact that the average costs of administration and non-purchased services is \$200 per client-month. Given that the average service spell length is 6 quarters, these costs are estimated to be \$3600 ($= 3 * 6 * 200$) per client. These two estimates reflect our uncertainty about the costs of non-purchased services and administration. Cases of individuals with mental illness may differ from the general population in both average purchased service expenditures and average spell lengths. So, if cases of individuals with mental illness have low average purchased services relative to non-purchased service costs, the first approach would be downward-biased. If instead, such cases have relatively low average costs associated with administration or non-purchased services, the second approach would be upward-biased. Finally, note that we do not compute separate estimates based on client-specific information on purchased services and spell length. We choose to use only an average "fixed" cost because the model and estimation procedure used to infer benefits does not allow actual expenditures to affect labor market outcomes.

Comparing these estimated costs and benefits reveals that DRS services provided to mentally ill have a substantial positive return especially in the longer run. In total, mean benefits range from \$2614 for the short run to \$5728 for the long run including *diagnosis & evaluation* and \$7256 for the short run to \$14283 for the long run excluding *diagnosis & evaluation*, while mean costs range from \$3200 to \$5000. Thus, whether there is a net positive or negative return depends on the approach used to infer costs associated with non-purchased services and administration and how one interprets estimates associated with *diagnosis & evaluation*. However, even under the most conservative assumptions, the long-run social benefit is estimated to exceed cost by 14%.

We also can compute the rate of return for each person receiving services in our sample. The results of this exercise are reported in Figure 14. For each sample individual receiving some service, we compare the expected flow of benefits they would get with the service package they received relative to the flow of benefits they would get with no services. We approximate cost as[45]

$$f + \sum_{j=1}^{J} y_{ij} c_j$$

where f is a combination of administrative costs and average (unobserved) in-house service and similar benefits costs, y_{ij} is an indicator for receipt of service j by person i (as defined in equation 1), and c_j is the average cost associated with service j computed as the ratio of "mean expenditure" and "% with positive expenditure" in Table 19. Figure 14 shows the distribution of quarterly rates of return for six scenarios: three with $f = \$1800$ and three with $f = \$3600$; and, for each assumption about f, we consider a) a 10-year horizon excluding *diagnosis & evaluation*, b) a 10-year horizon including *diagnosis & evaluation*, and c) a 5-year horizon excluding *diagnosis & evaluation*.[46] First, it is clear that earnings flows in years 6 through 10 have a significant impact on estimated rates of return, at least for conventional rates of return.[47] Thus, it is important to use long panels of earnings data such as ours when estimating rates of return.[48] Also it is clear that exclusion of observations receiving only *diagnosis & evaluation*

[45] An alternative is to use actual cost for each individual. The attractive feature of such an approach is that there is significant variation in cost even conditional on the set of services received. However, we choose to use only average costs for each service because, in the model and estimation procedure, we do not allow actual expenditures to affect labor market outcomes.

[46] Note that, when excluding *diagnosis & evaluation*, we a) ignore observations receiving only *diagnosis & evaluation* and b) ignore all costs and benefits associated with receipt of *diagnosis & evaluation*.

[47] At very high rates of return, later years become irrelevant because of the implied heavy discounting. For example, at a 20% quarterly rate of return, the discount factor associated with earnings 6 years in the future is 0.013.

[48] Estimated rates for returns for non-VR government training programs aimed at economically disadvantaged people also tend to be sensitive to short versus long horizons, and vary widely across programs, demographics, and studies. In some cases, these training programs are found to have average rates of return that are negative. But, in many others, the average annual rates of return are in excess of 100% (Friedlander, 1997; and LaLonde, 1995).

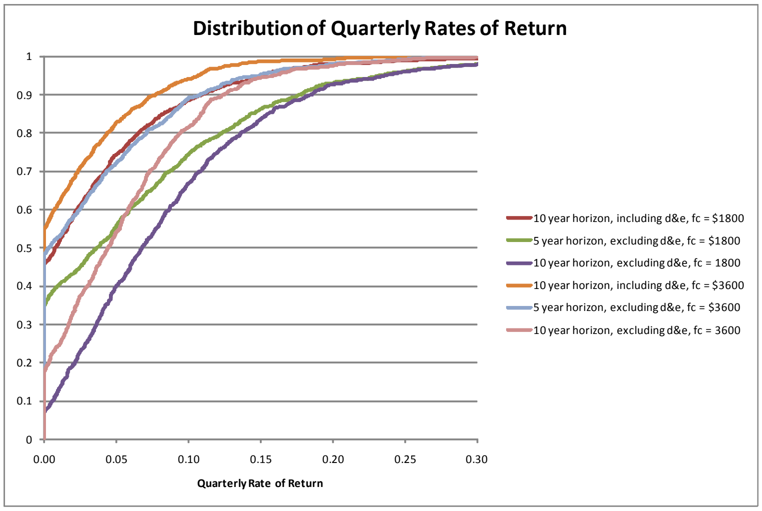

Figure 14: Distribution of Quarterly Rates of Return

has a significant impact on the distribution. See the discussion around Figure 11 on how one might interpret results for purchased *diagnosis* & *evaluation* services. Focusing on the distribution curve associated with a 10-year horizon and excluding *diagnosis* & *evaluation*, one sees that 6.9% of clients with mental illness have negative rates of return if $f = \$1800$ and 17.6% have negative rates of return if $f = \$3600$ (i.e., there is no positive discount rate that will justify the cost of services relative to the flow of future benefits). At the same time, even if $f = \$3600$, the median rate of return is quite high at 4.4% quarterly (18.9% annually), and 10% of rates of return are above 12.5% quarterly (60.1% annually); if $f = \$1800$, the median rate of return is 6.8% quarterly (30.1% annually), and 10% of rates of return are above 18.5% quarterly (97.2% annually). Meanwhile, including *diagnosis* & *evaluation* in the analysis causes the proportion with negative returns to increase significantly (for $f = \$3600$, it increases from 17.6% to 54.5%). Likewise, the proportion with negative returns increases significantly when focusing on the distribution curves associated with a 5-year horizon. It should be noted that the variation in rates of return here are due solely to variation in observable characteristics of individuals and variation in the set of services they receive; it is not due to randomness inherent in labor market experience.

Recently, there have been a number of state-level return on investment evaluations of VR services produced by economic consulting firms or university research bureaus (e.g., Heminway and Rohani, 1999; Uvin, Karaaslani, and White, 2004; Hollenbeck and Huang, 2006; Kisker et al., 2008; and Wilhelm and Robinson, 2010).[49] By comparing outcomes of a "treated" and "untreated"

[49] These state level studies condition the analysis on observed covariates. In some cases (Hollenbeck and Huang, 2006), researchers use statistical matching estimators based on propensity

group, as we do in Figures 5 and 6, these studies tend to find large positive returns to VR services. An evaluation of Utah's VR program, for example, found that the public benefits of the program, measured in dollars, exceed the cost by a factor of 5.64 (Wilhelm and Robinson, 2010). These reports, however, have a number of serious shortcomings which are addressed in this paper, including a) identification problems; b) problems caused by censored data; c) the selection problem; and d) heterogeneity in the caseload and in the services provided. Our analysis of the Virginia VR program addresses important limitations of these recent studies. First, using the model described in Section 2, we formally account for the possibility that selection into the treatment is endogenous. As noted above, a simple comparison of mean outcomes among treated and untreated clients may be spurious due to selection, and conditioning on observed covariates is not likely to address this problem credibly. Our results suggest that selection plays an important role in inferences on the effect of VR services. Second, by focusing on clients with mental illnesses, we allow the estimated effects of treatment to vary with the clients' limiting conditions. In contrast, these state-level reports do not distinguish between clients with mental illness, cognitive impairments, and physical impairments. Arguably, the effects of the program are heterogeneous, and restricting the impact to be constant across all groups may lead to biased inferences. Third, unlike these earlier evaluations, we examine the impact of specific types of services rather than just a single treatment indicator. We find that services do, in fact, have very different impacts on labor market effects. Finally, we observe labor market outcomes many years before and after the provision of VR services. In this analysis, being able to estimate the long-run return is critical as it significantly differs from the short-run return on investment.

7 Conclusions

Our results suggest a complex picture of the impact of VR services on labor market outcomes. Pre-program labor market differences vary across the six service types, estimated employment effects are positive for some services (e.g., *training*) and negative for others (e.g., *education*), and estimated earnings effects are consistently positive. When combining the employment and earnings effects together, we find that, except for *diagnosis & evaluation*, all of the other service types have positive long-run effects. On average, *training*, *restoration*, and *other services* have average benefits on the order of \$8000, while *education* and *maintenance* have positive average benefits of about \$2000. Overall, we find that VR services have a positive average return, with mean long-run benefits of \$5700 or \$14000, depending upon how one interprets *diagnosis & evaluation* results, and mean costs between \$3200 and \$5000. We also find,

scores, initially developed by Rosenbaum and Rubin (1983) and incorporated in other manpower training program evaluations (e.g., Heckman, Ichimura, and Todd,1997; Dehejia and Wahba, 1999). All of these analyses, however, invoke a conditional independence assumption that the outcome is independent of provision of services.

however, much variation in the return across VR participants. Depending upon how one estimates fixed costs (and excluding *diagnosis* & *evaluation*), between 6.9% (f = \$1800) and 17.6% (f = \$3600) of VR participants with mental illness have negative long-run rates of return, half have long-run rates of return in excess of between 30.1% (f = \$1800) and 18.9% (f = \$3600) annually, and 10% have annual long-run rates of return in excess of between 97.2% (f = \$1800) and 60.1% (f = \$3600).

8 Appendix

8.1 Covariance Structure

The covariance matrix of the errors $u_i' = (u_{i1}^y, u_{i2}^y, .., u_{iJ}^y, u_{i1}^z, u_{i1}^w, .., u_{iT}^z, u_{iT}^w)$ implied by the structure in equation (4) is

$$\Omega = \begin{pmatrix} A & B' \\ B & C + D \end{pmatrix} \tag{12}$$

where

$$A = \begin{pmatrix} \sum_k (\lambda_{1k}^y)^2 & \sum_k \lambda_{1k}^y \lambda_{2k}^y & \cdots & \sum_k \lambda_{1k}^y \lambda_{Jk}^y \\ \sum_k \lambda_{1k}^y \lambda_{2k}^y & \sum_k (\lambda_{2k}^y)^2 & \cdots & \sum_k \lambda_{2k}^y \lambda_{Jk}^y \\ \vdots & \vdots & \ddots & \vdots \\ \sum_k \lambda_{1k}^y \lambda_{Jk}^y & \sum_k \lambda_{2k}^y \lambda_{Jk}^y & \cdots & \sum_k (\lambda_{Jk}^y)^2 \end{pmatrix},$$

$$\begin{aligned} C &= H \otimes \underbrace{\begin{pmatrix} 1 & 1 & \cdots & 1 \\ 1 & 1 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & 1 \end{pmatrix}}_{T \times T} \\ H &= \begin{pmatrix} \sum_k (\lambda_k^z)^2 & \sum_k \lambda_k^z \lambda_k^w \\ \sum_k \lambda_k^z \lambda_k^w & \sum_k (\lambda_k^w)^2 \end{pmatrix}, \end{aligned}$$

$$D = \frac{\sigma_\zeta^2}{1 - \rho_\eta^2} \begin{pmatrix} 1 & \rho_\zeta & \rho_\eta & \rho_\eta \rho_\zeta & \cdots & \rho_\eta^{T-1} & \rho_\eta^{T-1} \rho_\zeta \\ \rho_\zeta & 1 & \rho_\eta \rho_\zeta & \rho_\eta & \cdots & \rho_\eta^{T-2} \rho_\zeta & \rho_\eta^{T-1} \\ \rho_\eta & \rho_\eta \rho_\zeta & 1 & \rho_\zeta & \cdots & \rho_\eta^{T-2} & \rho_\eta^{T-2} \rho_\zeta \\ \rho_\eta \rho_\zeta & \rho_\eta & \rho_\zeta & 1 & \cdots & \rho_\eta^{T-2} \rho_\zeta & \rho_\eta^{T-2} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \rho_\eta^{T-1} & \rho_\eta^{T-1} \rho_\zeta & \rho_\eta^{T-2} & \rho_\eta^{T-2} \rho_\zeta & \cdots & 1 & \rho_\zeta \\ \rho_\eta^{T-1} \rho_\zeta & \rho_\eta^{T-1} & \rho_\eta^{T-2} \rho_\zeta & \rho_\eta^{T-2} & \cdots & \rho_\zeta & 1 \end{pmatrix},$$

and

$$B = \begin{pmatrix} \sum_k \lambda_{1k}^y \lambda_{1k}^z & \sum_k \lambda_{2k}^y \lambda_{1k}^z & \cdots & \sum_k \lambda_{Jk}^y \lambda_{1k}^z \\ \sum_k \lambda_{1k}^y \lambda_{1k}^w & \sum_k \lambda_{2k}^y \lambda_{1k}^w & \cdots & \sum_k \lambda_{Jk}^y \lambda_{1k}^w \\ \vdots & \vdots & \ddots & \vdots \\ \sum_k \lambda_{1k}^y \lambda_{1k}^z & \sum_k \lambda_{2k}^y \lambda_{1k}^z & \cdots & \sum_k \lambda_{Jk}^y \lambda_{1k}^z \\ \sum_k \lambda_{1k}^y \lambda_{1k}^w & \sum_k \lambda_{2k}^y \lambda_{1k}^w & \cdots & \sum_k \lambda_{Jk}^y \lambda_{1k}^w \end{pmatrix}.$$

8.2 Counselor and Field Office Effects

We use as an instrument in equation (1), a transformation of the proportion of other clients of the same counselor provided service j, i.e., a counselor effect. We also use a transformation of the proportion of other clients from the same office provided service j, i.e., an office effect. We transform the counselor and office effects using an inverse normal distribution function to make it more likely that, as the counselor and office effects vary, their effect on service probabilities can vary by approximately the same amount. To consider why this is attractive, consider a counselor who almost always uses a particular service. We want to allow for the possibility that this will imply that all of the clients of the counselor are very likely to receive that service. Limiting the counselor effects to vary between $(0, 1)$ makes it harder for that to occur. On the other hand, using an inverse distribution function for a distribution with the real line as support makes the range $(-\infty, \infty)$.

While such a transformation makes sense analytically, in practice, it might cause problems for values of the untransformed effect at or near the boundaries. We propose a "fix" that both makes sense and solves the boundary problem. In particular, we propose replacing the untransformed effect r_{ij} with

$$r^*_{ij} = (1 - \omega_i) r_{ij} + \omega_i \overline{r}_j \quad (13)$$

where $\overline{r}_j$ is the mean value of r_{ij} across all counselors (offices), $\omega_i = \kappa_i^{-1}$, and κ_i is the number of clients seen by counselor i (office i). This specification allows the counselor effect and office effect to be more important for those counselors (offices) who have many observed clients. In fact, it has a certain Bayesian flavor to it.

There are some respondents who either have missing counselor or office information or who have a counselor (or office) with no other clients. For such cases, we can not create our effects.[50] Because of such cases, we include a set of dummies for missing counselor and/or missing office effects. It turns out that these dummies are very highly correlated, and most of the missing office effects must be excluded from the model to avoid a singular Hessian.

Tables A.1 and A.2 provide information about the moments of the transformed counselor and office effects. One can see that there is significant variation in both. There is some evidence of left-tailed skewness but no unreasonable outliers. The lack of outliers occurs despite zeroes for some services for some counselors and field offices because of the weighted average inherent in equation (13).

8.3 Local Labor Market Conditions

Virginia is unique among states in that it has both counties and independent cities. While BEA provides data for almost all counties and independent cities, there is a small number of mostly rural counties for which BEA provides data

[50] In fact, when a counselor (office) has only one other client, we treat it as missing also.

Table A.1: Moments of Inverse Normal Transformed Office Effects

Service	Mean	Std Dev	Minimum	Maximum
Diagnosis & Evaluation	-0.396	0.264	-1.512	0.785
Training	-0.149	0.255	-1.133	0.707
Education	-1.200	0.388	-2.699	-0.189
Restoration	-0.719	0.463	-2.469	0.511
Maintenance	-0.502	0.410	-1.754	0.419
Other Service	-0.828	0.595	-2.668	0.405

Note: # Obs = 1489.

Table A.2: Moments of Normal Logistic Transformed Counselor Effects

Service	Mean	Std Dev	Minimum	Maximum
Diagnosis & Evaluation	-0.412	0.424	-2.061	1.045
Training	-0.173	0.513	-1.795	1.472
Education	-1.351	0.625	-2.542	0.66
Restoration	-0.805	0.615	-2.298	0.735
Maintenance	-0.549	0.564	-2.105	0.802
Other Service	-0.883	0.697	-2.303	1.054

Note: # Obs = 1485.

Table A.3: Aggregated Regions

Region	Component Counties	Component Independent Cities
Eastern Shore	Accomack, Northhampton	
Rural Shenandoah, South	Bath, Highland, Rockbridge	Alleghany, Covington, Buena Vista, Lexington
Dinwiddie	Brunswick, Lunenberg, Nottoway	Dinwiddie, Colonial Heights, Petersburg, Greensville, Emporia
Bluefield, WV-VA Micropolitan SA	Bland, Buchanan, Dickerson, Lee, Norton, Smyth, Tazewell, Wythe, Wise	
Lynchburg Rural	Buckingham, Prince Edward	
Danville Rural	Charlotte, Halifax, Mecklenburg	
Northern Neck	Essex, King George, Lancaster, Middlesex, Northumberland, Richmond, Westmoreland	
Martinsville Rural	Floyd, Grayson, Patrick	Carroll, Galax
Culpeper, VA Micropolitan SA	Culpeper, Madison, Orange, Rappahannock	
Franklin/ Southhampton	Franklin, Southhampton	
Harrisonburg Rural	Page, Shenandoah	

only after some aggregation. We create 11 aggregated regions to deal with this problem listed in Table A.3.

We construct employment rate by dividing number of people employed by working age population. We do this both at the county/independent city level and at the MSA level. Significant variation in these measures exists across time, across geography, and across the two separate measures. One should note that there are some counties with employment rates greater than one. This occurs because the population numbers are based on county of residence while the employment numbers are based on county where one works. Thus, these rates reflect variation in net commuting patterns across counties.

9 References

References

[1] Aakvik, A., J. Heckman, and E. Vytlacil (2005). "Estimating Treatment Effects For Discrete Outcomes When Responses To Treatment Vary: An

Application to Norwegian Vocational Rehabilitation Programs." *Journal of Econometrics*. 125: 15-51.

[2] Abbrevaya, Jason, and Stephen Donald (2011). "A GMM Approach for Dealing with Missing Data on Regressors and Instruments." Unpublished manuscript.

[3] Arrighi, Y., B. Davin, A. Trannoy, and B. Ventelou (2010). "Un Effet Prix dans la Demande d'APA." Unpublished manuscript.

[4] Ashenfelter, Orley (1978). "Estimating the Effect of Training Programs on Earnings." *Review of Economics and Statistics*. 60: 47-57.

[5] Baldwin, Marjorie (1999). "The Effects of Impairments on Employment and Wages: Estimates from the 1984 and 1990 SIPP." *Behavioral Sciences and the Law*. 17: 7-27.

[6] Baldwin, Marjorie (2005). "Explaining the Differences in Employment Outcomes Between Persons With and Without Mental Disorders." Unpublished manuscript.

[7] Bassi, L. (1984). "Estimating the Effect of Training with Non Random Selection." *Review of Economics and Statistics*. 66: 36-43.

[8] Bell, S., L. Orr, J. Blomquist, and G. Cain (1995). *Program Applicants as a Comparison Group in Evaluating Training Programs*. Upjohn Institute for Employment Research.

[9] Bellante, D. (1972). "A Multivariate Analysis of a Vocational Rehabilitation Program." *Journal of Human Resources*. 7: 226-241.

[10] Berkowitz, E. (1988). "The Cost-Benefit Tradition in Vocational Rehabilitation." *Measuring the Efficiency of Public Programs: Costs and Benefits in Vocational Rehabilitation*. (ed.) E. Berkowitz. Temple University Press, 10-28.

[11] Bond, G., R. Drake, and D. Becker (2001). "An Update on Randomized Controlled Trials of Evidence-Based Supported Employment." *Psychiatric Rehabilitation Journal*. 31(4): 280-290.

[12] Börsch-Supan, Axel and Vassilis Hajivassiliou (1992). "Health, Children, and Elderly Living Arrangements: A Multiperiod-Multinomial Probit Model with Unobserved Heterogeneity and Autocorrelated Errors." *Topics in the Economics of Aging*. (ed.) David Wise. Chicago and London: U. of Chicago Press, 79-104.

[13] Bover, O., M. Arellano, and S. Bentolila (2002). "Unemployment Duration, Benefit Duration and the Business Cycle." *Economic Journal*. 112(479): 223–265.

[14] Bratberg, E., A. Grasdal, and A. Risa. (2002). "Evaluating Social Policy by Experimental and Nonexperimental Methods." *Scandinavian Journal of Economics*. 104(1): 147-171.

[15] Bureau of Economic Analysis (2010a). http://www.bea.gov/regional/docs/footnotes.cfm?tablename=CA04.

[16] Bureau of Economic Analysis (2010b). http://www.bea.gov/regional/docs/footnotes.cfm?tablename=CA25.

[17] Burns, Tom, Jocelyn Catty, Thomas Becker, Robert Drake, Angelo Fioritti, Martin Knapp, Christoph Lauber, Wulf Rössler, Toma Tomov, Jooske van Busschbach, Sarah White, Durk Wiersma, Connie Geyer, Marsha Koletsi, Pascale Lissouba, Miles Rinaldi, Ulrike Ehiosun, Rana Kalkan, Reinhold Kilian, Denise Manchisi, Astrid Niersman, Ingeborg Warnke, Dimitar Germanov, Adelina Comas, Claire Curran, and Anita Patel (2007). "The Effectiveness of Supported Employment for People with Severe Mental Illness: A Randomized Controlled Trial." *Lancet*. 370(9593): 1146-1152.

[18] Butler, Richard, William Johnson, and Marjorie Baldwin (1995). "Managing Work Disability: Why First Return to Works is Not a Measure of Success." *Industrial and Labor Relations Review*. 48(3): 452-469.

[19] Campbell, Kikuko, Gary Bond, Robert Drake, Gregory McHugo, and Haiyi Xie (2010). "Client Predictors of Employment Outcomes in High-Fidelity Supported Employment: A Regression Analysis." *Journal of Nervous and Mental Disease*. 198(8): 556-563.

[20] Cimera, R. (2010). "The National Cost-Efficiency of Supported Employees with Intellectual Disabilities: 2002 to 2007." *American Journal on Intellectual and Developmental Disabilities*. 115: 19-29.

[21] Clapp, Christopher, Amalia Miller, John Pepper, and Steven Stern (2010). "Evaluation of Programs for High Risk Youth in Virginia." Unpublished manuscript.

[22] Conley, R. (1969). "A Benefit Cost Analysis of the Vocational Rehabilitation Program." *Journal of Human Resources*. 4: 226-252.

[23] Cook, Judith, Anthony Lehman, Robert Drake, William McFarlane, Paul Gold, H. Stephen Leff, Crystal Blyler, Marcia Toprac, Lisa Razzano, Jane Burke-Miller, Laura Blankertz, Michael Shafer, Susan Pickett-Schenk, and Dennis Grey (2005). "Integration of Psychiatric and Vocational Services: A Multisite Randomized, Controlled Trial of Supported Employment." *American Journal of Psychiatry*. 162(10): 1948-1956.

[24] Currie, Janet and Bridget Madrian (1999). "Health, Health Insurance, and the Labor Market." *Handbook of Labor Economics, Volume 3*. (eds.) Orley Ashenfelter and David Card. Amsterdam : Elsevier.

[25] Dean, David and Robert Dolan (1991). "Fixed-Effects Estimates of Earnings Impacts for the Vocational Rehabilitation Program" *Journal of Human Resources*. 26(2): 380-391.

[26] Dean, David, Robert Dolan, Robert Schmidt, Paul Wehman, John Kregel, and Grant Revell (2002). "A Paradigm for Evaluation of the Federal-State Vocational Rehabilitation Program." *Achievements and Challenges in Employment Services for People with Disabilities: The Longitudinal Impact of Workplace Supports*. (eds.) John Kregel, David Dean, and Paul Wehman. Richmond: Virginia Commonwealth University Rehabilitation Research and Training Center on Workplace Supports.

[27] Dean, David, John Pepper, Robert Schmidt, and Steven Stern (2010a). "The Effects of Vocational Rehabilitation for People with Cognitive Impairments." Unpublished manuscript.

[28] Dean, David, John Pepper, Robert Schmidt, and Steven Stern (2010b). "The Effects of Youth Transition Programs on Labor Market Outcomes." Unpublished manuscript.

[29] Dehejia, R. and S. Wahba (1999). "Causal Effects in Non-Experimental Studies: Reevaluating the Evaluation of Training Programs." *Journal of the American Statistical Association*. 94(448): 1053-1062.

[30] Doyle, Joseph (2007). "Child Protection and Child Outcomes: Measuring the Effects of Foster Care." *American Economic Review*. 97(5): 1583-1610.

[31] Ettner, Susan, Richard Frank, and Ronald Kessler (1997). "The Impact of Psychiatric Disorders on Labor Market Outcomes." *Industrial and Labor Relations Review*. 51(1): 64-81.

[32] Farber, Henry (2010). "Job Loss and Unemployment in the 21st Century: The Great Recession in Labor Market Perspective." Unpublished Manuscript.

[33] Friedlander, Daniel, David Greenberg, and Philip Robins (1997). "Evaluating Government Training Programs for the Economically Disadvantaged." *Journal of Economic Literature*. 35(4), 1809-1855.

[34] Frolich, M., A. Heshmati, and M. Lechner (2004). "A Microeconometric Evaluation of Rehabilitation of Long-Term Sickness in Sweden." *Journal of Econometrics*. 19: 375–396.

[35] Gangl, Markus (2003). *Unemployment Dynamics in The United States and West Germany: Economics Restructuring, Institutions and Labor Market Processes*. Physica-Verlag Heidelberg, Germany, Chapter 4.

[36] Geweke, John (1988). "Antithetic Acceleration of Monte Carlo Integration in Bayesian Inference." *Journal of Econometrics*. 38(1/2): 73-89.

[37] Governmental Accountability Office (2005) "Vocational Rehabilitation: Better Measures and Monitoring Could Improve the Performance of the VR Program." Washington, D.C.

[38] Hall, Robert (2005). "Job Finding, and Unemployment in the U.S. Economy over the Past Fifty Years." *NBER Macroeconomics Annual*. 20: 101-137.

[39] Heckman, J. and V. Hotz (1989). "Choosing Among Alternative Nonexperimental Methods for Estimating the Impact of Social Programs: The Case of Manpower Training." *Journal of the American Statistical Association*. 84: 862-874.

[40] Heckman, J., H. Ichimura, and P. Todd (1997). "Matching as an Econometric Estimator: Evidence from Evaluating a Job Training Program." *Review of Economic Studies*. 65: 261-294.

[41] Heckman, James, Robert LaLonde, and Jeffrey Smith (1999). "The Economics and Econometrics of Active Labor Market Programs." *Handbook of Labor Economics Volume 3A*. (eds.) Orley Ashenfelter and David Card. Amsterdam: North-Holland. 1865-2097.

[42] Heckman, James and Burton Singer (1984a). "Econometric Duration Analysis." *Journal of Econometrics*. 24: 63-132.

[43] Heckman, James and Burton Singer (1984b). "A Method for Minimising the Impact of Distributional Assumptions in Econometric Models for Duration Data." *Econometrica*. 52(2): 271-320.

[44] Heminway, Derek and Faranak Rohani (1999). *A Cost-Benefit Analysis of the Employment of People with Disabilities in Florida: Final Report*. Educational Services Program.

[45] Hollenbeck, Kevin and Wei-Jang Huang (2006). *Net Impact and Benefit-Cost Estimates of the Workforce Development System in Washington State*. Kalamazoo, MI: W.E. Upjohn Institute for Employment Research.

[46] Hotz, V. (1992). "Design Analysis Plan for the GAO Study of the Vocational Rehabilitation Program." Correspondence with the GAO Program Evaluation and Methodology Division, 1-10.

[47] Imbens, Guido and Jeffrey Wooldridge (2009). "Recent Developments in the Econometrics of Program Evaluation." *Journal of Economic Literature*. 47(1): 5–86.

[48] Keith, Samuel, Darrel Regier, and Donald Rae (1991). *Psychiatric Disorders in America: The Epidemiologic Catchment Area Study*. (eds.) Lee Robins and Darrel Regier. MacMillan: New York.

[49] Kessler, R., P. Berglund, M. Bruce, J. Koch, E. Laska, P. Leaf, R. Manderscheid, R. Rosenheck, E. Walters, and P. Wang (2001). "The Prevalence and Correlates of Untreated Serious Mental Illness." *Health Services Research*. 36(6 Pt 1): 987-1007.

[50] Kimmel, Jean and Thomas Kniesner (1998). "New Evidence on Labor Supply: Employment versus Hours Elasticities by Sex and Marital Status." *Journal of Monetary Economics.* 42: 289-301.

[51] Kisker, E., G. Strech, J. Vetter, and C. Foote (2008). *Evaluating Rehabilitation Services in Oklahoma: An Analysis of Program Impacts and of Benefit/Costs*. Oklahoma Department of Rehabilitative Services.

[52] Kregel, John and D. Dean (2002). "Sheltered vs. Supported Employment: A Direct Comparison of Long-Term Earnings Outcomes for Individuals with Cognitive Disabilities." *Achievements and Challenges in Employment Services for People with Disabilities: The Longitudinal Impacts of Workplace Supports*. (eds.) John Kregel, David Dean, and Paul Wehman. Richmond: Virginia Commonwealth University Rehabilitation Research and Training Center on Workplace Supports.

[53] LaLonde, Robert (1995). "The Promise of Public Sector-Sponsored Training Programs." *Journal of Economic Perspectives*. 9(2), 149-168.

[54] Loprest, P. (2007). *Strategic Assessment of the State of the Science in Research on Employment for Individuals with Disabilities*. National Institute for Disability and Rehabilitation Research. Contract GS23F8198H.

[55] Machin, S., and A. Manning (1999). "The Causes and Consequences of Long-Term Unemployment in Europe." *Handbook of Labor Economics, Volume 3C.* (eds.) Orley Ashenfelter and David Card. Amsterdam : Elsevier.

[56] Marcotte D., V. Wilxox-Gok, and D. Redmond (2000). "The Labor Market Effects of Mental Illness: The Case of Affective Disorders." *The Economics of Disability*. (eds.) D. Salkever and A. Sorkin. Greenwich, CT: JAI Press.

[57] McFadden, Daniel and Kenneth Train (2000). "Mixed MNL Models for Discrete Response." *Journal of Applied Econometrics*. 15(5): 447-470.

[58] McKeithen, Amelia and Steven Stern (2007). "The Costs and Benefits of Mental Healthcare."
http://www.people.virginia.edu/~sns5r/resint/psychtstf/mentalhealthbenefits.pdf.

[59] Meyer, Bruce (1990). "Unemployment Insurance and Unemployment Spells." *Econometrica*. 58(4): 757-782.

[60] Meyer, Bruce (1995). "Natural and Quasi-Experiments in Economics." *Journal of Business and Economic Statistics.* 13(2):151–162.

[61] Moffitt, Robert (1985). "Unemployment Insurance and the Distribution of Unemployment Spells." *Journal of Econometrics*. 28: 85-101.

[62] Mueser, K., M. Salyers, and P. Mueser (2001). "A Prospective Analysis of Work in Schizophrenia." *Schizophrenia Bulletin*. 27: 281-296.

[63] Mueser, Peter, Kenneth Troske, and Alexey Gorislavky (2007). "Using State Administrative Data to Measure Program Performance." *Review of Economics and Statistics*. 89(4): 761-783.

[64] Nowak, L. (1983). "A Cost Effectiveness Evaluation of the Federal/State Vocational Rehabilitation Program Using a Comparison Group." *The American Economist*. 27: 23-29.

[65] Raphael, Steven, and Lorien Rice (2002). "Car Ownership, Employment, and Earnings." *Journal of Urban Economics*. 52(1): 109-130.

[66] Raaum, O., and H. Torp (2001). "Labor Market Training in Norway – Effects on Earnings." *Labour Economics*. 8(5): 207-247.

[67] Rosenbaum, P. and D. Rubin (1983). "The Central Role of the Propensity Score in Observational Studies for Casual Effects." *Biometrika*. 70: 41-55.

[68] Shimer, Robert (2007). "Reassessing the Ins and Outs of Unemployment." NBER Working Paper, No. 13421.

[69] Stern, Steven (1992). "A Method for Smoothing Simulated Moments of Discrete Probabilities in Multinomial Probit Models." *Econometrica*. 60(4): 943-52.

[70] Stern, Steven (1997). "Simulation-Based Estimation." *Journal of Economic Literature*. 35(4): 2006-2039.

[71] Stern, Steven (2010). "Moments and Extrema of Estimated Marginal Effects for 'The Effects of Vocational Rehabilitation for People with Mental Illness."
http://www.people.virginia.edu/~sns5r/resint/wrkdisstf/deanstruct.html.

[72] U.S. Dept. of Health and Human Services, National Center for Health Statistics (2004). *National Health Interview Survey*.

[73] Uvin, J., D. Karaaslani, and G. White (2004). *Evaluation of Massachusetts' Public Vocational Rehabilitation Program: Final Report*. Massachusetts Rehabilitation Commission.

[74] Van Den Berg, G., and J. Van Ours (1996). "Unemployment Dynamics and Duration Dependence." *Journal of Labor Economics.* 14(1): 100–125.

[75] Wilhelm, S. and J. Robinson (2010). *Utah State Office of Rehabilitation Economic Impact Study*. The University of Utah Center for Public Policy and Administration.

[76] Worrall, John (1978). "A Benefit-Cost Analysis of the Vocational Rehabilitation Program." *Journal of Human Resources*. 13: 285-298.