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

September 2015

Abstract

We construct a structural model of participation in vocational rehabilitation for people with mental illness. There are multiple services to choose among, and each has different effects on employment, earnings, and receipt of DI/SSI. We estimate large effects for most of the services implying large rates of return to vocational rehabilitation.

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 engage in gainful employment and possibly reducing disability insurance roles (Loprest, 2007), very little is known about the long term-efficacy of VR in the United States. Nearly all of the economic evaluations of the U.S. public-sector VR programs were conducted over 15 years ago (e.g., Dean, Dolan, and Schmidt, 1999) and a series of recent GAO (2005; 2012) reports stress the critical need for credible impact evaluations of VR programs. This need for updated information is only intensified with the recent passage of the Workforce Innovation and Opportunity Act of 2014 which requires VR programs to report post-closure employment and earnings.

In this paper, we study the impact of the VR program on persons diagnosed with mental illness. 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 mental illness. More than one-quarter of U.S. adults are mentally ill (Kessler et. al., 2001), and these illnesses are associated with severe employment consequences with unemployment rates for persons with severe mental illness estimated to be as high as 95% (Mueser, Salyers, and Mueser, 2001). While comprising an ever-larger share of the VR clientele, this group has turned out to be particularly hard to serve. As the United State Government Accountability Office (GAO) (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 has been concentrated on increasing the employability of persons with mental health problems.

---

*We would like to thank Joe Ashley, John Phelps, Kirsten Rowe, Ann Stanfield, Vlad Mednikov, and Jim Rothrock from Virginia Department of Aging and Rehabilitation Services (DARS), Paul O’Leary of the Social Security Administration, David Stapleton, other members of the National Institute of Disability and Rehabilitation Research (NIDRR) Advisory Group, Jeff Smith, and workshop participants at the ASSA Meetings, the Summer Econometric Society Meetings, APPAM, Clemson, Western Ontario, Washington University, University of Southern California, University of Louisville, University of Zurich, Hong Kong University of Science and Technology, National University of Singapore, Singapore Management University, and the University of Maryland for excellent help and advice and Rachel Fowley for excellent research assistance. The DARS, 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.

1One exception is Dean et. al. (2015a). 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. Early evaluations include Conley (1969); Bellante (1972); Worrall (1978); Berkowitz (1988); and Dean and Dolan (1991).

2Many earlier evaluations of VR programs do not distinguish between the different impairments. Arguably, however, the effects of the program are heterogeneous, and restricting the impact to be constant across all groups may lead to biased inferences (Dean and Dolan, 1991; Baldwin 1999; Dean, Dolan, and Schmidt, 1999; and Marcotte, Wilcox-Gok, and Redmond, 2000, Dean et. al., 2015a).
Using a unique panel data source on all persons who applied for services from the Department of Aging and Rehabilitative Services (DARS) in the state of Virginia in State Fiscal Year (SFY) 2000, we are able to estimate the long-term impact of VR services on employment, earnings, and disability insurance receipt. These results then are used to simulate the distribution of rates-of-return on VR services. Importantly, these administrative data from the 2000 applicant cohort in Virginia are much richer than that used in previous analyses. Most notably, we observe quarterly employment and earnings data as well as VR service data from 1995 to 2008. Observing individual quarterly employment and earnings prior to, during, and after service receipt, we examine both the short- and long-term effects of VR services.\(^3\) In addition, unlike much of the previous research, we examine specific types of services rather than just a single treatment indicator. Following Dean et al. (2002), we aggregate VR services into six types – diagnosis and evaluation, training, education, restoration, maintenance, and other services – and allow these six service types to have different labor market effects.

Another important contribution afforded by the richness of our data is that we evaluate the impact of VR services on the receipt of payments from the Social Security Administration’s (SSA) Disability Insurance (DI) and Supplemental Security Income (SSI) programs. As the enrollment and costs of disability insurance programs have grown over the past two decades, there has been growing interest in whether VR programs might serve to reduce the number of persons receiving DI/SSI benefits (e.g., Autor and Duggan, 2010; Stapleton and Marin, 2012). This is especially true for persons with mental illness who constitute the largest and most rapidly expanding subgroup of DI/SSI program beneficiaries (Drake et al., 2009). If VR services improve labor market outcomes of potential DI/SSI beneficiaries, some clients may choose to fully participate in the labor market rather than take up DI/SSI. Yet, VR programs may instead lead to an increase in take-up by serving to help clients understand the DI/SSI programs and rules (Stapleton and Martin, 2012).\(^4\)

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 labor market outcomes. 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). 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 that accounts for unobserved heterogeneity that may affect VR service receipt and labor market outcomes.

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 and augment that with a probit-like employment equation, an earnings equation, and a DI/SSI receipt equation. We allow for correlation of errors among all of the equations. This model has a rich structure of heterogeneous VR service receipt but a simplified structure for DI/SSI receipt. In particular, while the two SSA programs are quite different, with different earnings eligibility criteria as well as different benefits and benefit reduction rates, the model only indicates whether the VR applicant received either DI and/or SSI. While this simplified structure cannot measure how the specific features of the DI and SSI programs interact with the labor market and VR, it does allow us to estimate how VR impacts DI/SSI receipt. This is the first paper to jointly estimate VR service receipt, employment outcomes, and DI/SSI receipt. Moreover, Dean et. al. (2014) provide some support for this simplified model when they find nearly identical negative associations between VR receipt and SSI and DI receipt. Adding a detailed and nuanced model of DI and SSI receipt might be useful but would substantially complicate the model and would require data on disability status and labor market histories that are not included in the DARS data.\(^5\) For example, the DARS data do not reveal whether applicants are medically eligible for DI/SSI receipt.

After presenting the model, we then describe the four sources of data used in our analysis in Section 3 and the econometric methodology in Section 4. Estimation results and specification tests are presented in Section 5. Importantly, we develop an innovative test for whether the decision to apply for VR services is exogenous. Our analysis focus on program applicants, some of whom receive substantial services and others do not. Using this new test, we accept the null hypothesis that the decision to apply for VR services is exogenous. Finally, a rate-of-return analysis is presented in Section 6. Our results imply

---

\(^3\)Earlier economic analyses of VR efficacy relied almost exclusively on the RSA-911 Case Service Report of nationwide closures from the VR program. At the time, the RSA-911 form provides self-reported weekly earnings only at two points: 1) 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.

\(^4\)There are several studies assessing the correlation between VR services and DI/SSI receipt (e.g., Rogers, Bishop, and Crystal, 2005; and Stapleton and Erickson, 2004), but research on the impact of VR on DI/SSI receipt is limited (Stapleton and Martin, 2012).

\(^5\)As discussed in Gilleskie and Hoffman (2014), the complexities involved in modeling the dynamic interactions between DI, SSI and the labor market, along with the limitations of many datasets including the one used for this study, make it difficult to fully model these SSA programs. As a result, only a handful of papers jointly explore the SSI and DI application and award process, and the interaction of DI and SSI on labor market outcomes. See, for example, Kreider (1998, 1999), Benitez-Silva et al., (1999), Maestas et. al., (2013), and French and Song (2014).
generally high rates of return but with significant variation in returns across people with varying characteristics. We also find that VR services increase the probability of DI/SSI receipt.

2 Model

We use multivariate discrete choice model for service provision choices, and augment that with a probit-like employment equation, an earnings equation, and a DI/SSI receipt equation. As in much of the program evaluation literature, we focus on program applicants and do not model the application decision. In Section 5.4 we test for whether the decision to apply for services is exogenous.

Let $y_{ij}^*$ be the value for individual $i$ of participating in VR service $j$, $j = 1, 2, \ldots, J$, and define $y_{ij} = 1 (y_{ij}^* > 0)$ be an indicator for whether $i$ receives service $j$.\(^6\) Assume that

$$y_{ij}^* = X^y_{ij} \beta_j + u_{ij}^y + \varepsilon_{ij},$$

where $X^y_{ij}$ is a vector of exogenous explanatory variables and $u_{ij}^y$ is an error whose structure is specified below.

Next, we introduce three equations associated with the value of working, log-quarterly earnings, and the value of receiving DI/SSI payments. Let $z_{it}^*$ be the value to $i$ of working in quarter $t$, and define $z_{it} = 1 (z_{it}^* > 0)$). Assume that

$$z_{it}^* = X^z_{it} \gamma + \sum_{k=1}^K d_{ik} \sum_{j=1}^J \alpha_{jk}^z y_{ij} + u_{it}^z + v_{it}^z,$$

where $X^z_{it}$ 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 $\tau_k$ and $\tau_{k+1}$, 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 onset, b) 1 quarter before service onset, c) 1 quarter after service onset to 8 quarters after service onset, and d) 9 or more quarters after service onset. Let $w_{it}$ be the log quarterly earnings of $i$ at $t$, and assume that

$$w_{it} = X^w_{it} \delta + \sum_{k=1}^K d_{ik} \sum_{j=1}^J \alpha_{jk}^w y_{ij} + u_{it}^w + v_{it}^w,$$

where variables are defined analogously to equation (2). Let $r_{it}^*$ be the value to $i$ of receiving SSI or DI payments in quarter $t$, and define $r_{it} = 1 (r_{it}^* > 0)$. Assume that

$$r_{it}^* = X^r_{it} \psi + \sum_{k=1}^K d_{ik} \sum_{j=1}^J \alpha_{jk}^r y_{ij} + u_{it}^r + v_{it}^r,$$

where variables also are defined analogously to equation (2).\(^7\)

\(^6\)The issue of who is the decision-maker is important clinically. There are three possible decision-makers: the individual applying for service, a family member of the individual, and the DARS counselor. The goal is for all three decision-makers to work together to construct a plan that is best for the applicant. We have no information in our data on how decisions were actually made; nor do we know how one could measure that. Deviations from optimality may occur either because of the applicant’s mental illness or because the family member and/or the DARS counselor may have a different objective function than the applicant.

\(^7\)As noted above, the specification in equation (4) ignores the issues associated with actually applying for and being awarded disability benefits (e.g., see Kreider, 1998, 1999; Benitez-Silva et al., 1999; French and Song, 2014) or controlling for measurement error in disability and its interaction with disability benefits (e.g., see Benitez-Silva et al., 1999).
Finally, to allow for a rich correlation structure within and across these equations, assume that

\begin{align}
  u_{ij}^y &= \lambda_{ij}^y e_{i1} + \lambda_{ij}^z 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, \\
  u_{it}^r &= \lambda_1^r e_{i1} + \lambda_2^r e_{i2} + \eta_{it}^r, \\
  \eta_{it}^z &= \rho_{0} \eta_{it-1} + \zeta_{it}^z, \\
  \eta_{it}^w &= \rho_{0} \eta_{it-1} + \zeta_{it}^w, \\
  \eta_{it}^r &= \rho_{0} \eta_{it-1} + \zeta_{it}^r, \\
  \begin{pmatrix}
    \zeta_{it}^z \\
    \zeta_{it}^w \\
    \zeta_{it}^r
  \end{pmatrix} &\sim iidN \left[0, \Omega_{\zeta}\right], \\
  \begin{pmatrix}
    e_{i1} \\
    e_{i2}
  \end{pmatrix} &\sim iidN \left[0, I\right], \\
  v_{it}^z &\sim iidN \left[0, 1\right], \\
  v_{it}^w &\sim iidN \left[0, \sigma_w^2\right], \text{ and} \\
  v_{it}^r &\sim iidN \left[0, 1\right].
\end{align}

The norm in much of this literature is to specify person-specific unobserved heterogeneity as a one-factor model with a normal density, a one-factor model with a multi-point discrete distribution, or a two-factor model with a mixture of normal random variables for each factor (e.g., Heckman, Stixrud, and Urzúa, 2006; Conti, Heckman, and Urzúa, 2010). We include the \((e_{i1}, e_{i2})\) to allow for two common factors affecting all dependent variables with factor loadings \(\left(\lambda_{jk}^z, \lambda_k^w, \lambda_k^r\right)_{k=1}^s\). We also allow for serial correlation and contemporaneous correlation in the labor market errors \((\eta_{it}^z, \eta_{it}^w, \eta_{it}^r)\). The covariance matrix implied by this error structure is presented in Appendix 8.1. See Dean et al. (2015a) for a similar structure applied to people with cognitive impairments.

3 Data

We use three main sources of data: a) the administrative records for the state fiscal year (SFY) 2000 applicant cohort of DARS, b) the quarterly administrative records on labor market activity of the Virginia Employment Commission (VEC) from 1995 to 2008 for those people in the DARS data, and c) the quarterly administrative records of SSDI and SSI benefit receipt from 1995 to 2008 for those people in the DARS 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 DARS Data

3.1.1 DARS Sample Frame

Our starting point is the administrative records of the Virginia DARS 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 DARS clients with mental illnesses. There are two primary reasons why applicants are dropped from the sample. First, we exclude 6476 individuals who do not have 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. Second, we exclude 1220 applicants 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, 1984). 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

---

8Heckman, Stixrud, and Urzúa (2006) and Conti, Heckman, and Urzúa (2010) are more general than our specification, but they rely on the existence of other information about the latent factors to identify the mixture. The models with one factor and a multi-point discrete distribution are more general than our specification (given the asymptotic approximation described in Heckman and Singer, 1984) in the number of parameters included in the unobserved heterogeneity density but less general in terms of the number of independent factors to use.
Table 1: Proportion Receiving DRS Purchased and Non-Purchased Service by Type

<table>
<thead>
<tr>
<th>Service Type</th>
<th>Purchased Services</th>
<th>Non-Purchased Services Only</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Total w/ Positive Expenditures</td>
<td>w/ Zero Expenditures</td>
</tr>
<tr>
<td>Diagnosis &amp; Evaluation</td>
<td>0.547</td>
<td>0.477</td>
</tr>
<tr>
<td>Training</td>
<td>0.372</td>
<td>0.362</td>
</tr>
<tr>
<td>Education</td>
<td>0.132</td>
<td>0.131</td>
</tr>
<tr>
<td>Restoration</td>
<td>0.319</td>
<td>0.318</td>
</tr>
<tr>
<td>Maintenance</td>
<td>0.301</td>
<td>0.300</td>
</tr>
<tr>
<td>Other Service</td>
<td>0.234</td>
<td>0.232</td>
</tr>
</tbody>
</table>

first spell is in SFY 2000. Dean et al. (2015a) find significant left-censoring biases for a sample of people with cognitive impairments.9

3.1.2 DARS 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 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. There are 76 separate services provided by DARS and other state agencies. Following Dean et al. (2002), we aggregate these services into the six service types listed in Table 1. As discussed above, diagnosis & evaluation10 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 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 psychological services, dental services, hearing/speech services, eyeglasses, drug and alcohol treatments, surgical procedures, hospitalization, prosthetic devices, and other assistive devices.11 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.

The DARS administrative records provide access to types of purchased services, and we measure non-purchased from other reports from DARS (see Appendix (8.2)). The first column of Table 1 shows the proportion of the sample receiving purchased services, while the last column shows the prevalence in the sample for those receiving non-purchased services only.12 Although the vast majority of vouchers provide the cost of the purchase service, some provide no information about cost. Thus, in Table 1, total purchased services (column 1) are disaggregated into those with recorded positive expenditures and those without (columns 2 and 3).

Diagnostic and evaluation services are purchased in 54.7% of the cases, and non-purchased in additional 6.0% of the cases. Purchased services are provided in less than 40% of the cases for every other service type. This should be qualified

9 We also drop 701 applicants who were younger than 21 years in SFY 2000 and 59 who did not reside in Virginia. Finally, we exclude 94 applicants where primary and/or secondary diagnosis was missing, 71 with neither any service records nor employment records, and 147 applicants with other missing data problems.

10 We put variable names in a different font to avoid confusion.

11 For mentally ill people, psychological services are the most commonly utilized restorative service. Nearly half of all purchased restoration service vouchers are for psychological services and these services are provided to over three-quarters of the clients receiving restoration. The next largest subcategory (medical services) is only provided to 15.8% of the clients receiving restoration, and other subcategories are not provided at all (e.g., vision).

12 If the only source of service receipt is non-purchased, then the $\beta_j$ coefficients in equation (1) are multiplied by a non-purchased service-choice parameter (to be estimated), and the $\left(\alpha_{jk}^x, \alpha_{jk}^y, \alpha_{jk}^w\right)$ coefficients in equations (2), (3), and (4) are multiplied by an outcomes non-purchased service parameter (to be estimated). This allows both the service choice decisions, labor market, and DI/SSI outcomes to depend upon the source of the service. If a particular service type is both purchased and non-purchased, we report the individual as receiving the service only once.
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.

3.1.3 DARS Data for Explanatory Variables

Table 2 provides the sample moments for the explanatory variables coming from the DARS data. Variables indicate race and gender, whether the respondent is married (10.3%), received government assistance (19.1%), had a drivers licence (67.8%) and/or access to transportation (74.1%), and the number of dependents, (# dependents). Special education is a dummy variable equal to 1 for those observations where the respondent received a special education certificate; 2.5% of the respondents received such education. Education information is missing for 10.3% of the sample. Rather than exclude such observations, we include a dummy variable for when education information was missing. The government assistance variable measures the amount of assistance the applicant received from TANF, SSI, SSDI, or other government programs. On average, applicants received $191 in the application quarter.\footnote{13}

In addition, we also include a number of indicators of physical and mental disabilities. 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.\footnote{14} An individual’s counselor also assesses the significance of the disability. 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.\footnote{15}

Finally, to identify the impact of services on labor market outcomes and DI/SSI receipt, we use two instrumental variables that are correlated with the treatment assignment but not included in the labor market equations (2) and (3) or the DI/SSI receipt equation (4). 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 Appendix 8.3. Doyle (2007), Maestas, Mullen, and Strand (2013), and Dean et al. (2015a, 2015b) use similar instruments.

Two features of these instrumental variables are worth highlighting. First, the number of clients varies across counselors and field offices. For example, 43% of counselors have caseloads 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. Some applicants have missing counselor or office information, or have a counselor (or office) with one or two clients. For such cases, we include a variable indicating the counselor information is missing.\footnote{16}

\begin{table}[h]
\centering
\begin{tabular}{|l|c|c|l|c|c|}
\hline
\textbf{Variable} & \textbf{Socio-Demographic Variables} & \textbf{Disability Variables} & \textbf{Mean} & \textbf{Std Dev} & \textbf{Mean} & \textbf{Std Dev} \\
\hline
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 & 0.300 & Substance Abuse Problem & 0.151 & 0.358 \\
Age (Quarters/100) & 1.427 & 0.407 & \\
Married & 0.178 & 0.383 & Extent & 0.393 & 0.446 \\
# 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 & \\
\hline
\end{tabular}
\caption{Moments of Explanatory Variables}
\end{table}

\footnote{14}While some variables such as marital status, access to transportation, and number of dependents may be endogenous, we follow the literature (e.g., Ettner, Frank, and Kessler, 1997) and include them as significant indicators of inclusion in society and responsibility. Likewise, the variable measuring government financial assistance may be endogenous. However, we do formally model the DI/SSI receipt, and Gilleskie et al. (2014) find that, after controlling for a rich set of indicators of participation in different government assistance programs, benefit amounts may be exogenous. Finally, for this population, the income thresholds at which government benefits are reduced or eliminated is relatively high.

\footnote{15}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.

\footnote{16}The meaning of mental illness as an explanatory variable is that, at the time of application, the individual’s primary or secondary diagnosis

\footnote{17}Of the 1555 applicants, there are 70 with missing counselor variables and 66 with missing office variables. Missing office instruments are highly

\end{document}
Table 3: Moments of Employment and Earnings Variables

<table>
<thead>
<tr>
<th>Variable</th>
<th>Before Initial Service Quarter</th>
<th>After Initial Service Quarter</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td># Obs</td>
<td>Mean</td>
</tr>
<tr>
<td>Employment</td>
<td>31427</td>
<td>0.35</td>
</tr>
<tr>
<td>Log Quarterly Earnings</td>
<td>11003</td>
<td>7.082</td>
</tr>
</tbody>
</table>

Figure 1: Employment Rates

Second, the proportion of clients receiving each service varies by counselors and field offices. For example, 10% of offices provide diagnosis & evaluation services to less than 20% of their clients while 4% provide it for all of their clients. Overall, there is strong evidence of meaningful variation in behavior across counselors and field offices. Using a likelihood ratio test, we reject the null hypothesis that the joint density of services is the same across offices and that each office provides each service in the same proportion. Likewise, we reject these same nulls applied to counselors.\(^{18}\) The fact that there is significant variation in the provision of services across offices and counselors make our instrument viable. Whether these instruments satisfy other identification restrictions is evaluated in Section 4.2.

3.2 VEC Data

In addition to the DARS data, we also have data on individual quarterly earnings prior to, during, and after service receipt. In particular, this study uses quarterly employment records provided by employers to the Virginia Employment Commission (VEC) for purposes of determining eligibility for unemployment insurance benefits. Details on how the VEC data are matched to the DARS data are provided in Appendix 8.4.

In our analysis, we try to explain two labor market outcome variables: employment and log quarterly earnings.\(^{19}\) 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 decompose quarterly earnings into wage level and hours, this is not possible in the VEC data. Table 3 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).

Figures 1 and 2 display quarterly employment rates and earnings (conditional on employment), respectively, for SFY 2000 applicants who receive substantial 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.

correlated with the missing counselor variables, and are thus excluded from the model.

\(^{18}\) The test statistic for the office nulls are 407.44 (with 245 df and normalized value of 7.33) and 575.39 (with 294 df and a normalized value of 11.60), while for counselors the analogous test statistics are 970.60 (with 785 df and a normalized value of 4.68) and 3636.94 (with 942 df and a normalized value of 66.70).

\(^{19}\) 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.
Perhaps the most striking finding is seen in Figure 1 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 for the nine-years post application.

While there is notable association between DARS services receipt and employment, there is no such relationship with earnings. Figure 2 shows that quarterly earnings among the employed are almost identical for the treated and the untreated throughout. Thus, the data imply 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.

These results may suggest that VR programs effectively increase employment, but we caution against drawing this type of causal conclusion from this evidence alone. The observed post-application increase in employment rates for treated clients may be due to VR services, but it may also reflect unobserved heterogeneity associated with selection into treatment and labor market outcomes. This selection problem will be addressed using the structural model developed in Section 2.

### 3.3 SSA Data

A Memorandum of Agreement (MOA) between the SSA and DARS allowed us to obtain monthly SSI (Supplemental Security Income) and DI (Disability Insurance) payments for individuals in our cohort. We aggregate SSI and DI benefit recipiency into a single binary indicator of receipt of disability benefits from either source. The majority of DI/SSI recipients who apply for VR services are not employed. In our sample, we observe receipt of disability benefits in 38.4% of the person/quarters, with nearly four-fifths of these being during periods when the individual is not in the labor market. The correlation of disability benefit receipt and work activity (measured as a binary variable) is −0.293, and the person-specific correlation is −0.189.

Among DI/SSI recipients who are employed, earnings are modest (less than $300/qtr, on average), and the distribution decreases with earnings for both DI and SSI recipients. Interestingly, earnings do not appear to be sensitive to the threshold where one would start losing benefits (SGA). In particular, very few DI or SSI recipients appear to treat the SGA as a binding constraint; i.e., there is no “parking” right below the SGA (see Dean et al., 2014).

Still, as noted above, there may be important interactions between DI/SSI and labor market outcomes. While this analysis focuses on the impact of VR on DI/SSI receipt, other research has considered the complicated interactions between the SSA disability programs and the labor market. See, for example, Kreider (1999), Benitez-Silva et al. (1999), Kreider and Riphahn (2000), Maestas et al., (2013), and French and Song (2014).

### 3.4 BEA Data

Labor market outcomes may be influenced by local conditions. To account for local labor market differences, we construct measures of log employment rates using county level data from the Bureau of Economic Analysis (BEA) on population size and number of people employed, disaggregated by age and county (BEA, 2010a). These measures are merged to the VR
client data using geographic identifiers that allow us to match each DARS client with their county of residence. Details are included in Appendix 8.5.

### 4 Econometric Methodology

#### 4.1 Likelihood Function

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

\[
\theta_y = (\beta_j, \lambda_y^y, \lambda_y^w, 1)^T, \\
\theta_z = (\gamma, \lambda_z^1, \lambda_z^2, \rho_{\gamma}^z, \alpha_{jk})_{j=1}^J, \\
\theta_w = (\delta, \lambda_w^1, \lambda_w^2, \rho_{\gamma}^w, \sigma_w^2, \alpha_{jk})_{j=1}^J, \\
\theta_r = (\psi, \lambda_r^1, \lambda_r^2, \rho_{\gamma}^r, \alpha_{rk})_{k=1}^K.
\]

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

\[
L_i = \int L_i (u_i) \, dG (u_i | \Omega)
\]

where

\[
L_i (u_i) = L_i^0 (u_i^y) \prod_{t=1}^T L_{it}^{zw} (u_{it}^z, u_{it}^w, u_{it}^r),
\]

\[
L_i^0 (u_i^y) = \prod_{j=1}^J \frac{\exp \{ X_i^y \beta_j + u_j^y \}}{1 + \exp \{ X_i^y \beta_j + u_j^y \}},
\]

\[
L_{it}^{zw} (u_{it}^z, u_{it}^w, u_{it}^r) = \left[ L_{it}^0 (u_{it}^z, u_{it}^w) \right]^{1 - x_{it}} \left[ L_{it}^0 (u_{it}^z, u_{it}^w) \right]^{x_{it}} L_{it}^2 (u_{it}^r),
\]

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

\[
L_{it}^1 (u_{it}^z, u_{it}^w) = \frac{1}{\sigma_w} \phi \left( \frac{u_{it}^w - X_{it}^w \delta - \sum_{k=1}^K d_{ik} \sum_{j=1}^J \alpha_{jk} y_{ij} - u_{it}^z}{\sigma_w} \right),
\]

\[
\phi \left( X_{it}^z \gamma + \sum_{k=1}^K d_{ik} \sum_{j=1}^J \alpha_{jk} y_{ij} + u_{it}^z \right),
\]

\[
L_{it}^2 (u_{it}^r) = \Phi \left( X_{it}^r \psi + \sum_{k=1}^K d_{ik} \sum_{j=1}^J \alpha_{jk} y_{ij} + u_{it}^r \right)^{r_{it}}
\]

\[
\left[ 1 - \Phi \left( X_{it}^r \psi + \sum_{k=1}^K d_{ik} \sum_{j=1}^J \alpha_{jk} y_{ij} + u_{it}^r \right) \right]^{1 - r_{it}},
\]

and \( G (u_i | \Omega) \) is the joint normal density with covariance matrix \( \Omega \) described in Appendix 8.1. It is straightforward to simulate the integral in (6) using well-known methods described in Stern (1997). The functional form of the conditional likelihood contribution associated with observed program choices, \( L_{it}^0 (u_i^y) \) in equation (7), 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 and DI/SSI receipt, \( L_{it}^{zw} (u_{it}^z, u_{it}^w, u_{it}^r) \) in equations (8), (9), (10), and (11) follow from the normality
assumption for \((v_{it}^w, v_{it}^w, v_{it}^r)\) and the trivariate normality assumption for \((\zeta_{it}^v, \zeta_{it}^w, \zeta_{it}^r)\) in equation (5). The log likelihood function is \(L = \sum_{t=1}^{n} \log L_t\).\(^{20}\)

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, labor market outcome variables, and DI/SSI receipt 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 \(\beta_j\) coefficient in equation (1) associated with the male for \(j = \text{training}\). Similarly, the covariation between white and employment status identifies the \(\gamma\) coefficient in equation (2) associated with white, and the covariation between white and log quarterly earnings identifies the \(\delta\) coefficient in equation (3) associated with white. Similar sample covariances identify parameters for DI/SSI recipiency in equation (4). Second moment parameters such as \(\Omega_\xi\) in equation (5) are identified by corresponding second sample moments.

Two approaches are used to address the second identification problem. First, 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), (3) and (4), \(u_{it}\), are fixed over time, then controls for the observed pre-treatment labor market and DI/SSI receipt differences address the endogenous selection problem. 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).

Second, we include two instruments in equation (1) that are excluded from equations (2), (3) and (4). As described in Sections 3.1.3 and Appendix 8.3, 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 \(j\). Similar instruments are applied in Doyle (2007), Maestas, Mullen, and Strand (2013), and Dean et al. (2015a, 2015b).

Excluding these instruments from the labor market and DI/SSI equations seems sensible and, as illustrated in Sections 3.1.3 and 5.2, these variables are strongly correlated with service receipt. However, it must also be the case that these variables are exogenous or unrelated to the structural errors. While one can never be certain this holds, there are good reasons to think it is a reasonable assumption especially given that we include in the analysis the client’s observed limitations, county-level employment rates, and pre-service labor market outcomes. Most notably, DARS clients have limited ability to select their field office or counselor; the field office is determined by the residential location of the client, and, conditional on observed limitations, counselors are randomly assigned.\(^{21}\)

So, unless clients relocate to take advantage of the practices of particular field offices, the assignment to offices and counselors is effectively random conditional on the observed limitations of clients. A threat to the validity of these instruments may arise if variation in the availability of jobs where training (or other DARS services) is productive might jointly affect labor market outcomes and the average behavior of counselors and field offices. Including measures of local labor market conditions directly in labor market outcome equations (2) and (3) should ameliorate this problem. Another concern arises if there is unobserved variation in the ability of counselors to match clients with jobs that affects both his/her decisions about what types service to offer clients and later success in the labor market. We assume that this type of confounding effect is not important. Finally, one might worry that variation in counselor/office behaviors impacts the decision to apply for services and thus result in endogenous application decisions. In Section 5.4, however, we find evidence the decision to apply is exogenous.

Our approach for addressing the endogenous selection of services represents a substantial advance over the existing literature where the past research generally relies on limited controls for pre-program earnings and assumes service participation is otherwise exogenous. Along with Aakvik, Heckman, and Vytalci (2005) and Dean et al. (2015a, 2015b), this is one of the first studies to identify the impact of VR services on labor market outcome using both a history of pre-program earnings and plausibly exogenous instrumental variables.

\(^{20}\)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, and Geweke (1988) shows that the simulation error occurring in simulation-based estimators for a significant class of models is of order \((1/n)\) when antithetic acceleration is used. We simulate all errors except for \(\eta\) and \(\varepsilon\) with antithetic acceleration (Geweke, 1988) 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).

\(^{21}\)Counselors are assigned by office policy that does not involve client choice. For example, some field offices assign counselors to balance caseload across counselors, some have counselors who specialize in mental illness, and some assign counselors by client locale.
<table>
<thead>
<tr>
<th>Variable</th>
<th>Two or More Quarters Prior to Service Participation</th>
<th>Quarter Prior to Service Participation</th>
<th>First 2 Years After Service Participation</th>
<th>More than 2 Years After Service Participation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diagnosis &amp; Evaluation</td>
<td>0.280 ** (0.008)</td>
<td>0.091 (0.081)</td>
<td>0.052 ** (0.014)</td>
<td>-0.182 ** (0.007)</td>
</tr>
<tr>
<td>Training</td>
<td>-0.361 ** (0.010)</td>
<td>-0.168 * (0.101)</td>
<td>0.270 ** (0.017)</td>
<td>0.180 ** (0.008)</td>
</tr>
<tr>
<td>Education</td>
<td>0.283 ** (0.014)</td>
<td>0.175 (0.134)</td>
<td>-0.036 (0.025)</td>
<td>0.170 ** (0.010)</td>
</tr>
<tr>
<td>Restoration</td>
<td>0.275 ** (0.010)</td>
<td>0.461 ** (0.092)</td>
<td>0.258 ** (0.018)</td>
<td>0.148 ** (0.008)</td>
</tr>
<tr>
<td>Maintenance</td>
<td>-0.217 ** (0.011)</td>
<td>-0.222 ** (0.113)</td>
<td>-0.163 ** (0.019)</td>
<td>-0.291 ** (0.009)</td>
</tr>
<tr>
<td>Other Services</td>
<td>0.156 ** (0.011)</td>
<td>-0.035 (0.122)</td>
<td>0.284 ** (0.020)</td>
<td>0.205 ** (0.009)</td>
</tr>
</tbody>
</table>

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.

5 Estimation Results

5.1 Estimates of Impact of VR Services

We begin by examining the estimated effect of services on labor market outcomes. Table 4 presents the estimates and associated standard errors for the effect of services on employment, and Table 5 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. 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).

The first two columns of Tables 4 and 5, which display estimates for the quarters prior to the initial service, provide 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 restoration in the log quarterly earnings equation. For training and maintenance, the estimates imply that those individuals provided training services have lower pre-treatment employment probabilities and quarterly earnings. For education and other services, the estimates imply that selection is positively associated with pre-service labor market outcomes – people with higher pre-treatment employment rates and earnings are more likely to be assigned to these services. 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 column. For example, as seen in Table 4, prior to service provision, the employment propensity for clients provided training services is 0.361 less than for clients that do not receive these services. In the two years after the start of service provision, it rises to 0.270, and then, in the longer run, it declines to 0.180. Thus, after accounting for selection into service, the long-term effect of training on those who were trained is 0.180 + 0.361 = 0.541. 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, education, and restoration decrease employment propensity. Maintenance increases employment propensity in the short run but decreases it in the long run.

Table 5 shows that restoration, maintenance, and other services increase conditional earnings, relative to earnings two or more quarters prior to service provision, in both the short and long run, and diagnosis & evaluation, training, and education.
Table 5: DARS Purchased Service Participation Effects on Log Quarterly Earnings

<table>
<thead>
<tr>
<th>Variable</th>
<th>Two or More Quarters Prior to Service Participation</th>
<th>Quarter Prior to Service Participation</th>
<th>First 2 Years After Service Participation</th>
<th>More than 2 Years After Service Participation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diagnosis &amp; Evaluation</td>
<td>-0.017 ** (0.013)</td>
<td>-0.349 ** (0.107)</td>
<td>-0.102 ** (0.025)</td>
<td>0.015 (0.011)</td>
</tr>
<tr>
<td>Training</td>
<td>-0.065 ** (0.017)</td>
<td>-0.086 (0.159)</td>
<td>-0.120 ** (0.031)</td>
<td>0.071 ** (0.014)</td>
</tr>
<tr>
<td>Education</td>
<td>0.096 ** (0.022)</td>
<td>0.027 (0.186)</td>
<td>0.011 (0.042)</td>
<td>0.242 ** (0.017)</td>
</tr>
<tr>
<td>Restoration</td>
<td>-0.028 * (0.015)</td>
<td>-0.116 (0.123)</td>
<td>0.064 ** (0.031)</td>
<td>0.178 ** (0.013)</td>
</tr>
<tr>
<td>Maintenance</td>
<td>-0.293 ** (0.018)</td>
<td>-0.027 (0.169)</td>
<td>-0.187 ** (0.033)</td>
<td>-0.076 ** (0.013)</td>
</tr>
<tr>
<td>Other Services</td>
<td>0.070 ** (0.017)</td>
<td>-0.199 (0.176)</td>
<td>0.154 ** (0.032)</td>
<td>0.216 ** (0.014)</td>
</tr>
</tbody>
</table>

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.

Figure 3: DARS Purchased Service Effects on Long-Term Discounted Benefits

decrease conditional earnings in the short run but increase them in the long run.\(^{22}\)

Because of the variation in effects over time and over labor market outcomes, it is difficult to infer the long-run benefits of each service. Accordingly, Figure 3 reports the mean present value for 10 years of earnings flows (measured in $10000) excluding service costs, a 95% confidence range, and the minimum and maximum present value of each service.\(^{23}\) 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 $7200, $3750, and $4800 respectively, while education and maintenance have positive benefits of $1700 and $2100 respectively. It should be noted that 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 3 shows that the long-run conditional earnings effects essentially offset the negative employment effects for present value calculations.\(^{24}\) One other notable feature of the discounted benefits calculations illustrated in Figure 3 is the high degree of variability across the caseload. The discounted benefits associated with training services, for example, range from $700 to nearly $22800.

For diagnosis & evaluation the estimates are negative. While it may be that these services have a negative long-run

\(^{22}\) Almost 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.

\(^{23}\) We use a quarterly discount factor of 0.95. 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. Because the distribution of benefits is highly skewed, the normal approximation is not appropriate. Instead, we report the 0.025 and 0.975 quantiles of the empirical distribution.

\(^{24}\) Restoration also has negative employment outcomes but positive results for conditional earnings. For these services, service receipt appears to raise earnings and reservation wages. However, it is not clear why reservation wages would rise faster than wages (which would be necessary for employment effects to be negative).
Table 6: DARS Purchased Service Participation Effects on DI/SSI Participation Propensity

<table>
<thead>
<tr>
<th>Variable</th>
<th>Two or More Quarters Prior to Service Participation</th>
<th>Quarter Prior to Service Participation</th>
<th>First 2 Years After Service Participation</th>
<th>More than 2 Years After Service Participation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Two or More Quarters Prior to Service Participation</td>
<td>Quarter Prior to Service Participation</td>
<td>First 2 Years After Service Participation</td>
<td>More than 2 Years After Service Participation</td>
</tr>
<tr>
<td>Diagnosis &amp; Evaluation</td>
<td>-0.562 ** (0.014)</td>
<td>-0.139 (0.151)</td>
<td>0.244 ** (0.019)</td>
<td>0.667 ** (0.008)</td>
</tr>
<tr>
<td>Training</td>
<td>0.124 ** (0.031)</td>
<td>0.574 ** (0.189)</td>
<td>0.592 ** (0.027)</td>
<td>0.423 ** (0.039)</td>
</tr>
<tr>
<td>Education</td>
<td>-0.176 ** (0.026)</td>
<td>0.383 (0.282)</td>
<td>0.471 ** (0.044)</td>
<td>0.345 ** (0.013)</td>
</tr>
<tr>
<td>Restoration</td>
<td>-0.691 ** (0.028)</td>
<td>-0.779 ** (0.230)</td>
<td>-0.699 ** (0.025)</td>
<td>-0.161 ** (0.009)</td>
</tr>
<tr>
<td>Maintenance</td>
<td>-0.179 ** (0.016)</td>
<td>-0.009 (0.020)</td>
<td>0.193 ** (0.028)</td>
<td>0.054 ** (0.010)</td>
</tr>
<tr>
<td>Other Services</td>
<td>-0.545 ** (0.018)</td>
<td>-0.318 (0.230)</td>
<td>-0.366 ** (0.031)</td>
<td>-0.101 ** (0.019)</td>
</tr>
</tbody>
</table>

Notes:
1. Estimates are effects on DI/SSI Participation Propensity.
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.

In this scenario, where purchased diagnosis & evaluation services are provided to applicants with relatively poor labor market prospects, one might expect DARS counselors to influence these clients to move in a different, more rewarding direction. In fact, we find evidence that the purchased diagnosis & evaluation services increase the receipt of DI/SSI. Table 6 presents the results for the effect of services on DI/SSI receipt. The estimates show that every service increases the probability of DI/SSI receipt. For example, the long-term effect of training on the DI/SSI receipt propensity is estimated to be $0.423 - 0.124 = 0.299$, which implies that training services increase the probability of DI/SSI receipt on average by 0.05. Likewise, diagnosis & evaluation increase the probability of DI/SSI receipt by 0.18, education by 0.08, restoration by 0.06, maintenance by 0.03, and other services by 0.01. The effect for diagnosis & evaluation is the largest which lends more credence to the argument about the effect of purchased diagnosis & evaluation on labor market outcomes.

Most previous evaluations 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 1 and 2 above, which display the mean employment and earnings outcomes respectively, show little pre-program differences, fairly substantial positive post-treatment employment associations, and almost no relationship between treatment and earnings. The 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; third, by incorporating DI/SSI as a choice variable; and finally, by using instrumental variables in a model with endogenous service provision. The results from the estimates effect, there are other plausible explanations. This may reflect the fact that these services are provided largely in-house, yet our data do not fully measure non-purchased services provided by counselors (see Appendix 8.2). So, while nearly every applicant receives some diagnosis & evaluation services, our data indicated that only 60.7% of applicants receive services – 54.7% purchased and 6.0% non-purchased (see Table 1). In addition to this measurement problem, purchased diagnosis & evaluation services may differ from other types of service in ways that may bias the resulting estimates. In particular, purchased diagnosis & evaluation services are more likely to be provided to clients with problems that make it difficult for them to succeed in the labor market; 79% of those receiving purchased services are examined by specialists. Thus, heterogeneity in the severity of mental illness may be an unobserved factor related to the provision of diagnosis & evaluation services and labor market outcomes. Under reasonable modeling restrictions, the instrumental variable estimate on diagnosis & evaluation will be negatively biased and the estimate on mental illness will be upward biased. See Appendix 8.6 for an illustration.
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, estimated earnings effects that are consistently positive in the long run, and a positive effect of VR on DI/SSI receipt.

5.2 Estimates of Counselor and Office Effects

Table 7 presents estimates of the counselor and office instruments on VR service receipt. 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 variable defined in Appendix 8.3; note that these effects are restricted to be the same across different services. The coefficients associated with the missing counselor variables are the effect on \( y_{ij} \) when the relevant counselor does not have enough other clients to compute a set of counselor effects. The estimates imply that the counselor and office effects have a large and statistically significant effect on service receipt.

5.3 Estimates of the Covariance Structure

Our model has a rich error covariance structure, as seen in equation (5). 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 8 demonstrate no statistically significant correlations between the errors associated with the provision of service types and the error associated with employment propensity. However, the correlation between the errors for employment propensity and log quarterly earnings is positive (0.556 and 0.436) and the correlation between the errors for employment and DI/SSI receipt are negative (0.556 and −0.225). This suggests that there is some unobserved personal characteristic, maybe ability, that increases employment probabilities and conditional earnings but decreases DI/SSI receipt probabilities. This is consistent with the idea that VR helps those unlikely to succeed in the labor market take-up DI/SSI.

Meanwhile, the factor loadings for Factor 2 imply negative correlation between the errors associated with log quarterly earnings and DI/SSI receipt (−0.116 and 1.652). This suggests another unobserved characteristic, perhaps some other component of ability, increasing log quarterly earnings and decreasing DI/SSI receipt but having no real impact on employment propensity. The estimates of the factor loadings for service provision imply that neither unobserved component has any meaningful effect on service provision except for maintenance which is positively correlated with DI/SSI receipt (0.197 and 1.652).

---

25 Appendix (8.7) presents and evaluates the estimates of the demographic characteristics on the propensity to use different services (\( y_{ij}^{*st} \) in equation (1)) and the effects of demographic,socioeconomic, and disability-related characteristics on the three labor market outcomes of interest (\( z_{it} \) in equation (2), \( w_{it} \) in equation (3), and \( r_{it} \) in equation (4)). For the most part, the observed characteristics do not have statistically significant effects on service receipt, while almost all of the coefficient estimates in the labor market equations are statistically significant with the expected signs.

26 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.
Table 8: Covariance Factor Loadings

<table>
<thead>
<tr>
<th>Variable</th>
<th>Factor 1</th>
<th></th>
<th>Factor 2</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Diagnosis &amp; Evaluation</td>
<td>-0.055</td>
<td>0.065</td>
<td>-0.032</td>
<td>0.069</td>
</tr>
<tr>
<td>Training</td>
<td>0.006</td>
<td>0.085</td>
<td>-0.065</td>
<td>0.083</td>
</tr>
<tr>
<td>Education</td>
<td>-0.064</td>
<td>0.109</td>
<td>-0.152</td>
<td>0.110</td>
</tr>
<tr>
<td>Restoration</td>
<td>-0.039</td>
<td>0.075</td>
<td>-0.143 *</td>
<td>0.083</td>
</tr>
<tr>
<td>Maintenance</td>
<td>-0.056</td>
<td>0.086</td>
<td>0.197 **</td>
<td>0.088</td>
</tr>
<tr>
<td>Other Services</td>
<td>0.022</td>
<td>0.088</td>
<td>-0.257</td>
<td></td>
</tr>
<tr>
<td>Employment</td>
<td>0.556 **</td>
<td>0.003</td>
<td>-0.002</td>
<td>0.003</td>
</tr>
<tr>
<td>Log Quarterly Earnings</td>
<td>0.436 **</td>
<td>0.004</td>
<td>-0.116 **</td>
<td>0.004</td>
</tr>
<tr>
<td>SSI/DI</td>
<td>-0.225 **</td>
<td>0.003</td>
<td>1.652 **</td>
<td>0.005</td>
</tr>
</tbody>
</table>

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.

The estimates of the other elements of the error structure are reported in Appendix (8.1). All of the estimated covariance terms are relatively small and dominated by the factor structure terms in Table 8. The estimate of the log earnings error $\sigma_w$ is quite large, implying that a standard deviation in quarterly earnings due to unobserved factors is on the order of $88546$. 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 $-0.2$ but does not estimate hours effects.

5.4 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. Then we construct the standard $\chi^2$ test statistic. 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.27 The test statistics are $\chi^2_{31} = 288.29$ for employment probabilities before service receipt and $\chi^2_{32} = 109.87$ for employment probabilities after service receipt. Both of these are highly significant, implying a poor fit. Figure 4 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 good job predicting employment probabilities up to a predicted probability of about 0.7, after which the fit worsens. Overall, while we are predicting employment probabilities reasonably well, there is some concern that, between 0.2 and 0.5, we are underestimating employment probabilities prior to VR service receipt and overestimating it after VR service receipt; this may have a large effect on our rate-of-return analysis. We address this issue in Section 6.

For DI/SSI receipt, the test statistics are $\chi^2_{35} = 368.8$ for probabilities before service receipt and $\chi^2_{38} = 1218.1$ for probabilities after service receipt. However, the curves analogous to Figure 4 (see Appendix 8.8) look extremely similar to those in Figure 4; they fit well until predicted probabilities are above 0.7 (where there many fewer observations).

Using a series of Lagrange Multiplier (LM) tests, we consider allowing for interactions among pairs of services in the labor market outcome equations. While the nonlinearity of the model creates some interactions, it may not be appropriate to rely strictly on the model structure. These LM tests, however, suggest that it is not important to allow for service interactions. Similarly, we test for interactions between demographic characteristics (male and white) and services but find no evidence of such interactions.

Finally, we test for whether the decision to apply for VR services is exogenous. Our analysis above follows the conventional approach in evaluation of job training programs of ignoring selection problems associated with the application process (see Imbens and Wooldridge, 2009; Heckman et. al., 1999), and instead focusing on the assignment of employment services among

---

27 Before service receipt includes the quarter before receipt, and after service receipt includes quarters in the first two years after receipt and the longer run.
applicants. Yet, focusing on applicants may create biases and limit the external validity of the estimates if the decision to apply for services is related to unobserved factors associated with labor market outcomes. For example, the field office and caseworker instruments may be invalid if the application decision is impacted by the behavior of the VR offices or caseworkers. In addition, the model estimates may not be informative about policies that impact the applicant pool.

To investigate this issue, we use results in Johnson et al. (2015) to estimate the number of mentally ill persons in the cities and counties in Virginia and, along with our administrative data on the number of VR clients with mental illness, county level estimates of the application probability. To address the endogeneity concern, we use information on local (non-VR) level services provided to the mentally ill as an instrumental variable. The idea is that local services are related to whether a person applies to DARS but unrelated to the provision of DARS services. Given this information, we can test for whether the application decision is exogenous with respect to VR service provisions, labor market outcome, and DI/SSI receipt. In all cases, we accept the null hypothesis that the application decision is exogenous at the 5% significance level (in general, the $p$-values exceed 0.20). Details of the test and the test results are provide in Appendix 8.9.

6 Rate of Return

The preceding analysis suggests that, except for purchased diagnosis & evaluation services, observed DARS services have long-run positive effects on labor market outcomes (see Figure 3). 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 DARS data on the costs of purchased services. There are, however, many services for which we do not have direct evidence on the associated costs. In particular, the costs of non-purchased services are not observed in the DARS data file. For these items, we present more speculative evidence.

We simulate the private labor market benefits to DARS clients using the structural model estimates summarized in Section 5. 29 In particular, we compute the mean present discounted value of the provided services relative to the value of receiving no services using both a 5- and 10-year post-treatment observation period for those individuals who received some service and using an annual discount factor of 0.987. The estimated mean discounted benefits are $1942 with a standard deviation of $3726 using the 5-year window and $4124 with a standard deviation of $7677 using a 10-year window.

28 This test is robust to specification biases that might arise if we were to instead incorporate an application equation directly into the model. Moreover, the test does not require micro level data necessary for estimating an application equation which are not readily available.

29 This simulation has a similar structure to the one used to compute marginal effects in Section 5.1 (see Figure 3). But here we compute the present discounted value of the actual treatments provided by DARS 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.
As noted in Section 5, the estimated long-run benefits associated diagnosis & evaluation services is likely to be downward biased. Thus, a more accurate estimate of the mean discounted benefits of VR services may be found by excluding diagnosis & evaluation services from the benefits computation. While the extent of this bias is unknown, the best reason to think that they may have a true negative effect on labor market outcomes is that they encourage those who are unlikely to succeed in the labor market to sign up for DI/SSI benefits which, from a social welfare point of view, should not be thought of as a negative benefit. Thus, setting the benefits of diagnosis & evaluation to zero will lead to a conservative and more accurate estimate of the long-run benefits. In this scenario, we estimate that the mean discounted benefits are $4233 with a standard deviation of $3678 using the 5-year window and $8374 with a standard deviation of $7744 using a 10-year window.\footnote{There are several reasons these estimated benefits may not reflect the true social benefits of VR services. First, the estimated benefits do not reflect the potential 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 costs associated with welfare programs and increased VR participant utility due to reduced welfare dependence (Lalonde, 1995), improved health status, and access to health care insurance. At the same time, the deadweight costs of taxation may change if welfare receipt and tax payments change. Likewise, we do not include the estimated positive impact of VR services on DI/SSI receipt.}

Next, we consider the costs of providing VR services. The mean costs of purchased services is $1500.\footnote{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 allows neither service duration nor actual expenditures to affect labor market outcomes.} To estimate the costs for non-purchased services, we use information on DARS spending by fiscal year as reported to the US Social Security Administration. These reports summarize information on aggregate administrative costs, DARS-provided counseling, guidance, and placement service costs, purchased service cost, other in-house costs, and the size of the caseload for each fiscal year. While there is some variation in the distribution of costs across years, in general, non-purchased service and administrative costs account for 55% 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 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 $1500 per client, fixed costs are estimated to be $1800 (≈ ($1500/0.45) − $1500) 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 costs.\footnote{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.}

Comparing these estimated costs and benefits reveals that DARS services provided to mentally ill people have a positive return especially in the longer run. In total, our preferred estimates, which exclude diagnosis & evaluation, imply that mean benefits range from $4233 for the short run to $8374 for the long run, while mean costs range from $3300 to $5100. Thus, even under the most conservative assumptions about the costs of services, the long-run social benefit is estimated to exceed costs by 64%. With diagnosis & evaluation included, the mean long run benefit of $4124 lies within the two cost estimates.

We also can compute the rate of return for each person receiving services. The results of this exercise are reported in Figure 5. 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

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

where \( f \) is a combination of administrative costs and average non-purchased service 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.”

Figure 5 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.\footnote{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 allows neither service duration nor actual expenditures to affect labor market outcomes.} Two general lessons emerge from this figure. 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. Thus, it is important to use long panels of earnings data such as ours when estimating rates of return. Second, it is clear that exclusion of diagnosis & evaluation has a significant impact on the distribution.

The figure also illustrates wide variation in the rates of return across individuals. Focusing on the distribution curve associated with a 10-year horizon and excluding diagnosis & evaluation, the distribution curve shows that 6.8% 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 (60.0% annually); if \( f = \$1800 \), the median rate of return is 6.8% quarterly (30.2% annually), and 10% of rates of return are above 18.5% quarterly (97.3% 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.

Earlier, in Section 5.4, Figure 4 showed that the employment probabilities are overestimated after VR service provisions and underestimated before service. Together, these imply that we might be overestimating the effect of VR services on employment rates. Consider, for example, the distribution of quarterly rates of return for the 10-year horizon without diagnosis & evaluation and with estimated fixed costs of \$1800 \). Adjusting all of the predicted employment probabilities by the bias reported in Figure 4 causes the proportion of people with negative rates of return to increase from 6.8% to 14.6%, and the median quarterly return to decrease from 6.8% to 5.7% (24.8% annual rate of return). However, this bias correction should be interpreted only as suggestive because it does not control for other sources of bias and because we have not computed the standard error of the bias correction.

### 7 Conclusions

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). By comparing outcomes of a “treated” and “untreated” group, as we do in Figures 1 and 2, 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. 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

---

34 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).
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, sensory impairments, or 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 (Dean and Dolan, 1991; Baldwin 1999; Dean, Dolan, and Schmidt, 1999; and Marcotte, Wilcox-Gok, and Redmond, 2000, Dean et al., 2015a). Third, unlike these earlier evaluations, we examine the impact of specific types of services rather than just a single treatment indicator. We find that these services do, in fact, have very different impacts on labor market effects. Finally, we observe labor market and disability insurance 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.

Our results suggest a complex picture of the impact of VR services on labor market outcomes and DI/SSI receipt. 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 benefits on the order of $7200, $3750, and $4800 respectively, while education and maintenance have positive benefits of $1700 and $2100 respectively. Overall, we find that VR services have a positive average return, with mean long-run benefits of $4124 or $8374, depending upon how one interprets diagnosis & evaluation results, and mean costs between $3300 and $5100. We also find, however, much variation in the return across VR participants. Depending upon how one estimates fixed costs $f$ (and excluding diagnosis & evaluation), between 6.8% ($f = $1800) and 17.6% ($f = $3600) of VR participants with mental illness have negative long-run rates of return, half have long-run annual rates of return in excess of between 30.2% ($f = $1800) and 18.9% ($f = $3600), and 10% have annual long-run annual rates of return in excess of between 97.3% ($f = $1800) and 60.0% ($f = $3600). Finally, our results suggests that VR programs are unlikely to reduce the burgeoning growth in DI/SSI roles. To the contrary, we find that these services increase the probability VR clients take-up DI/SSI.

8 Appendices
8.1 Covariance Structure

The covariance matrix of the errors $u_i = (u_i^{y}, u_i^{z}, u_i^{w}, u_i^{r})$ implied by the structure in equation (5) of the paper is

$$
\Omega_{(J+3T) \times (J+3T)} = \begin{pmatrix}
A & B^T \\
B & C + D
\end{pmatrix}
$$

where

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

$$
C = H \otimes Q_T
$$

$$
H = \begin{pmatrix}
\sum_k (\lambda_{1k}^w)^2 & \sum_k \lambda_{1k}^w \lambda_{2k}^w & \cdots & \sum_k \lambda_{1k}^w \lambda_{jk}^r \\
\sum_k \lambda_{1k}^w \lambda_{2k}^w & \sum_k (\lambda_{2k}^w)^2 & \cdots & \sum_k \lambda_{2k}^w \lambda_{jk}^r \\
\vdots & \vdots & \ddots & \vdots \\
\sum_k \lambda_{1k}^w \lambda_{jk}^r & \sum_k \lambda_{2k}^w \lambda_{jk}^r & \cdots & \sum_k (\lambda_{jk}^r)^2
\end{pmatrix},
$$

$$
Q_T = \begin{pmatrix}
1 & 1 & \cdots & 1 \\
1 & 1 & \cdots & 1 \\
\vdots & \vdots & \ddots & \vdots \\
1 & 1 & \cdots & 1
\end{pmatrix}
$$
### Table A.1: Other Covariance Terms

<table>
<thead>
<tr>
<th>Variable</th>
<th>Estimate</th>
<th>Std Err</th>
<th>Variable</th>
<th>Estimate</th>
<th>Std Err</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\text{Var}(\zeta_1)$</td>
<td>0.041 **</td>
<td>0.000</td>
<td>$\text{Var}(\zeta_3)$</td>
<td>0.008 **</td>
<td>0.004</td>
</tr>
<tr>
<td>$\text{Cov}(\zeta_1,\zeta_2)$</td>
<td>0.018 **</td>
<td>0.004</td>
<td>$\text{Cov}(\zeta_1,\zeta_3)$</td>
<td>-0.018 **</td>
<td>0.005</td>
</tr>
<tr>
<td></td>
<td>0.000</td>
<td>0.000</td>
<td>$\sigma_w$</td>
<td>1.281 **</td>
<td>0.003</td>
</tr>
</tbody>
</table>

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

$$
D = \Omega_3 \otimes \frac{1}{1 - \rho_\eta^2} \begin{pmatrix}
1 & \rho_\eta & \ldots & \rho_\eta^{T-1} \\
\rho_\eta & 1 & \ldots & \rho_\eta^{T-2} \\
\vdots & \vdots & \ddots & \vdots \\
\rho_\eta^{T-1} & \rho_\eta^{T-2} & \ldots & 1 
\end{pmatrix},
$$

and

$$
B = q_T \otimes F,
$$

$$
q_T = \begin{pmatrix}
1 \\
1 \\
\vdots \\
1
\end{pmatrix},
$$

$$
F = \begin{pmatrix}
\sum_k \lambda_{yk}^y \lambda_{yk}^y & \sum_k \lambda_{yk}^y \lambda_{yk}^y & \ldots & \sum_k \lambda_{yk}^y \lambda_{yk}^y \\
\sum_k \lambda_{yk}^y \lambda_{yk}^y & \sum_k \lambda_{yk}^y \lambda_{yk}^y & \ldots & \sum_k \lambda_{yk}^y \lambda_{yk}^y \\
\vdots & \vdots & \ddots & \vdots \\
\sum_k \lambda_{yk}^y \lambda_{yk}^y & \sum_k \lambda_{yk}^y \lambda_{yk}^y & \ldots & \sum_k \lambda_{yk}^y \lambda_{yk}^y
\end{pmatrix}.
$$

The estimates of the primitives, other than the factor loading estimates in Table 8, associated with the covariance matrix are reported in Table A.1.

### 8.2 Measuring Non-Purchased Services

Services can be provided to an individual in any combination of three ways: a) as a “purchased service” through an outside vendor using DARS funds, b) as a “similar benefit” purchased or provided by another governmental agency or not-for-profit organization with no charge to DARS, and/or c) internally by DARS personnel. The DARS administrative data do not, however, provide the same detailed information for in-house services or similar benefits. Instead, we measure non-purchased service provision using two additional sources of service information. First, DARS reports on the provision of similar benefits (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. Use of this information is complicated by several factors, the most important being that the two indicators included for each service category sometimes provide inconsistent information. We impose the condition that this source identifies the provision of similar benefits only if both indicators designate service provision. Second, we observe data on in-house benefits provisions from the Woodrow Wilson Rehabilitation Center (WWRC), a state agency that provides comprehensive, individualized services with an employment objective. The WWRC receives an annual block grant from DARS which it administers autonomously. When appropriate, DARS refers individuals to the WWRC for rehabilitative services. The WWRC provided us with service information for this type of in-house benefit. Because there may be some classification errors between in-house services and similar benefits, we identify them simply as “non-purchased services.” These two sources of information cover all non-purchased service expenses except for in-house counselor services.

Although purchased services and in-house services provided by WWRC map uniquely into the six service categories used in our analysis, 4 of the 22 categories used for the RSA-911 do not. For example, the RSA category diagnostic & treatment includes both the diagnosis & evaluation category as well as the restoration category. Using diagnostic & treatment as an example, 6 of the 75 DARS purchased service categories map into diagnosis & evaluation, and 14 map into restoration. For the individuals flagged by RSA codes as having received diagnosis & evaluation, we count the number of sample individuals who received a service in one or more of the 6 diagnosis & evaluation purchased service codes ($D$) and the number of sample individuals in one or more of the 14 restoration codes ($R$). We then assign a probability that an individual designated in the
RSA-911 file as receiving diagnosis & evaluation receives diagnosis & evaluation as \( 0.56 = D/(D + R) \) and restoration as \( 0.44 = R/(D + R) \).

### 8.3 Counselor and Field Office Effects

We use as an instrument in equation (1) of the paper, 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 \( c_{ij} \) with

\[
c^*_ij = (1 - \omega_i) c_{ij} + \omega_i \bar{c}_j
\]

where \( \bar{c}_j \) is the mean value of \( c_{ij} \) across all counselors (offices), \( \omega_i = \kappa_i^{-1} \), and \( \kappa_i \) 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. 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 the missing office effects must be excluded from the model to avoid a singular Hessian.

Tables A.2 and A.3 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 (12).

### 8.4 VEC Data Match

DARS provided the VEC with identifiers from the universe of 10323 applicants for DARS services in SFY 2000. The VEC returned to DARS 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.\(^{36}\) Table

---

\(^{35}\)In fact, when a counselor (office) has only one other client, we treat it as missing also.

\(^{36}\)This analysis was not limited to applicants with mental illness diagnoses.
Table A.3: Moments of Normal Logistic Transformed Counselor Effects

<table>
<thead>
<tr>
<th>Service</th>
<th>Mean</th>
<th>Std Dev</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diagnosis &amp; Evaluation</td>
<td>-0.412</td>
<td>0.424</td>
<td>-2.061</td>
<td>1.045</td>
</tr>
<tr>
<td>Training</td>
<td>-0.173</td>
<td>0.513</td>
<td>-1.795</td>
<td>1.472</td>
</tr>
<tr>
<td>Education</td>
<td>-1.351</td>
<td>0.625</td>
<td>-2.542</td>
<td>0.66</td>
</tr>
<tr>
<td>Restoration</td>
<td>-0.805</td>
<td>0.615</td>
<td>-2.298</td>
<td>0.735</td>
</tr>
<tr>
<td>Maintenance</td>
<td>-0.549</td>
<td>0.564</td>
<td>-2.105</td>
<td>0.802</td>
</tr>
<tr>
<td>Other Service</td>
<td>-0.883</td>
<td>0.697</td>
<td>-2.303</td>
<td>1.054</td>
</tr>
</tbody>
</table>

Note: # Obs = 1485.

Table A.4: Comparison between SSA and VEC Employment Records

<table>
<thead>
<tr>
<th></th>
<th>2001</th>
<th>2002</th>
</tr>
</thead>
<tbody>
<tr>
<td>Neither SSA nor VEC show earnings</td>
<td>31%</td>
<td>35%</td>
</tr>
<tr>
<td>SSA shows earnings, VEC does not</td>
<td>12%</td>
<td>12%</td>
</tr>
<tr>
<td>VEC shows earnings, SSA does not</td>
<td>1%</td>
<td>1%</td>
</tr>
<tr>
<td>Both SSA &amp; VEC show earnings</td>
<td>57%</td>
<td>52%</td>
</tr>
<tr>
<td>Mean SSA Earnings</td>
<td>$9,117</td>
<td>$9,859</td>
</tr>
<tr>
<td>Mean SSA - VEC Difference</td>
<td>$510</td>
<td>$616</td>
</tr>
</tbody>
</table>

A.4 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. Although formally accounting for these coverage errors is beyond the scope of this paper, the results in Table A.4 suggests that any resulting biases should be minimal for the earnings equations but may be more important for the employment regressions. Unfortunately, our agreement with the SSA did not allow us to investigate whether these errors varied by VR service receipt.

8.5 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 only after some aggregation. We create 11 aggregated regions to deal with this problem listed in Table A.5.

We construct the 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.

8.6 Bias Caused by Unobserved Heterogeneity in Measured Mental Illness

There are many possibilities to explain the results with respect to the diagnosis & evaluation effects, including, but not limited to, the possibility that (a) the instrument is correlated with the errors; (b) the estimated net effect of diagnosis & evaluation on long-run outcomes is negative is a statistical anomaly and one might reject the (one-sided) null hypothesis that all of the long-run outcomes are positive; and (c) extra diagnosis and evaluation requires much time for people with mental illness and thus slows down the rehabilitation process (note that the negative effect is due solely to employment effects). The problem with (a) is that the result is specific to diagnosis & evaluation, and it disappears for other disability groups (e.g., Dean, et al., 2015a). The problem with (b) is that the null hypothesis would be rejected. We have no information on (c).

37In the paper, we use only the county/independent city level because the two measures are very highly correlated.
The bias explanation we prefer, which is also confirmed by DARS counselors, is the explanation included in the text of the paper. The idea is that, for people with mental illness, receipt of purchased diagnosis & evaluation services is an indicator that the individual’s mental health problem is particularly difficult to deal with in a way unobserved in the DARS administrative data. This unobserved heterogeneity in mental illness is an error in measurement of an explanatory variable, and it causes diagnosis & evaluation to be correlated with the errors in the labor market outcomes equations. More explicitly, but in a simpler linear context, consider the model,

\[ y_i = X_i \beta + w_{1i} \alpha_1 + w_{2i} \alpha_2 + u_i \]

where \( y_i \) is an outcome variable of interest for observation \( i \), \( X_i \) is a vector of exogenous explanatory variables, \( w_{1i} \) is a potentially endogenous explanatory variable such as receipt of diagnosis & evaluation, and \( w_{2i} \) is another explanatory variable measured with error such as degree of mental illness with

\[ \text{plim} \left( n^{-1} \sum_i w_{1i} w_{2i} \right) > 0. \]

In particular, for simplicity, assume that

\[ w_{2i} \in \{0, 1, 2\}, \]

but \( w_{2i} \) is not observed, and, instead,

\[ x_i = 1 \left( w_{2i} > 0 \right) \]

is observed. Then the model to be estimated

\[ y_i = X_i b + w_{1i} a_1 + x_i a_2 + v_i. \]
Let $Z$ be a matrix of instruments. Then the asymptotic properties of the IV estimator are

\[
\text{plim} \left( \begin{bmatrix} \hat{b} \\ \hat{a}_1 \\ \hat{a}_2 \end{bmatrix} \right) = \text{plim} \left( \frac{Z'X/n}{Z'w_1/n} \right) \text{plim} \left( \frac{Z'w_1/n}{Z'x/n} \right) \text{plim} \left( \frac{Z'y/n}{Z'x/n} \right)
\]

\[
= \text{plim} \left( \frac{Z'X/n}{Z'w_1/n} \right) \left[ \text{plim} \left( \frac{Z'X/n}{Z'w_1/n} \right) \left( \frac{\beta}{\alpha_1} \right) + \frac{Z'u/n}{Z'w_1/n} \right]
\]

\[
= \text{plim} \left( \frac{Z'X/n}{Z'w_1/n} \right) \text{plim} \left( \frac{Z'X/n}{Z'w_2/n} \right) \left( \frac{\beta}{\alpha_1} \right) \neq \left( \frac{\beta}{\alpha_2} \right).
\]

Now, in the interest of making more progress, consider a special case where

\[w_{1i} = \gamma_0 + \gamma_1 w_{2i} + e_i.\]

Then

\[
\text{plim} \left( \begin{bmatrix} \hat{b} \\ \hat{a}_1 \\ \hat{a}_2 \end{bmatrix} \right) = \text{plim} \left( \frac{Z'X/n}{Z'[\gamma_0 1 + \gamma_1 w_2 + e]/n} \right) \left( \frac{Z'[1 (w_2 > 0)]/n}{Z'[\gamma_0 1 + \gamma_1 w_2 + e]/n} \right) \left( \frac{\beta}{\alpha_1} \right).
\]

Next, in the same spirit, assume that $\beta = 0$; i.e., there are no $X$’s (without this assumption, as is the case in any measurement error problem, the sample correlation of $X$ with $w_1$ and $w_2$ contaminates the analysis relative to the simpler case). Then

\[
\text{plim} \left( \begin{bmatrix} \hat{a}_1 \\ \hat{a}_2 \end{bmatrix} \right) = \text{plim} \left( \frac{z'_1 [\gamma_0 1 + \gamma_1 w_2 + e]/n}{z'_1 w_2/n} \text{plim} \left( \frac{z'_x/n}{z'_2 w_2/n} \frac{z'_2 [\gamma_0 1 + \gamma_1 w_2 + e]/n}{z'_2 x/n} \right) \left( \frac{\alpha_1}{\alpha_2} \right) \right)
\]

\[
= \text{plim} \left( \frac{z'_1 [\gamma_0 1 + \gamma_1 w_2 + e]/n}{z'_1 w_2/n} \text{plim} \left( \frac{z'_x/n}{z'_2 w_2/n} \frac{z'_2 [\gamma_0 1 + \gamma_1 w_2 + e]/n}{z'_2 x/n} \right) \left( \frac{\alpha_1}{\alpha_2} \right) \right).
\]

where

\[
A_{11} = \left( \frac{z'_2 x}{n} \right) \left( \frac{z'_1 [\gamma_0 1 + \gamma_1 w_2 + e]/n}{z'_1 w_2/n} \right) - \left( \frac{z'_1 x}{n} \right) \left( \frac{z'_2 [\gamma_0 1 + \gamma_1 w_2 + e]/n}{z'_2 w_2/n} \right),
\]

\[
A_{12} = \left[ \frac{z'_2 x}{n} - \left( \frac{z'_2 w_2}{n} \right) \right] \left( \frac{z'_2 [\gamma_0 1 + \gamma_1 w_2 + e]/n}{z'_2 w_2/n} \right),
\]

\[
A_{21} = \left[ \frac{z'_1 w_2}{n} - \left( \frac{z'_1 x}{n} \right) \right] \left( \frac{z'_2 [\gamma_0 1 + \gamma_1 w_2 + e]/n}{z'_2 w_2/n} \right),
\]

\[
A_{22} = \left( \frac{z'_2 w_2}{n} \right) \left( \frac{z'_1 [\gamma_0 1 + \gamma_1 w_2 + e]/n}{z'_2 w_2/n} \right) - \left( \frac{z'_1 x}{n} \right) \left( \frac{z'_2 [\gamma_0 1 + \gamma_1 w_2 + e]/n}{z'_2 w_2/n} \right),
\]

\[
D = \left[ \frac{z'_2 x}{n} \right] \left( \frac{z'_1 [\gamma_0 1 + \gamma_1 w_2 + e]/n}{z'_2 w_2/n} \right) - \left( \frac{z'_1 x}{n} \right) \left( \frac{z'_2 [\gamma_0 1 + \gamma_1 w_2 + e]/n}{z'_2 w_2/n} \right).
\]
Note that, in the case where \( x = w_2 \) (i.e., there is no measurement error),

\[
\text{plim} \left( \frac{\hat{a}_1}{\hat{a}_2} \right) = \left( \begin{array}{cc} 1 & 0 \\ 0 & 1 \end{array} \right) \left( \begin{array}{c} \alpha_1 \\ \alpha_2 \end{array} \right) = \left( \begin{array}{c} \alpha_1 \\ \alpha_2 \end{array} \right)
\]

In general,

\[
\text{plim} \left( \frac{\hat{a}_1}{\hat{a}_2} \right) = \left( \begin{array}{cc} 1 & \gamma_{12} \\ \eta_{21} & 1 \end{array} \right) \left( \begin{array}{c} \alpha_1 \\ \alpha_2 \end{array} \right)
\]

where

\[
\eta_{12} = \frac{\text{plim} \left[ \left( \frac{z_{11}^*}{n} \right) - \left( \frac{z_{11}^{w_2}}{n} \right) \right] \left( \frac{z_{11}^*}{n} \right) \left( \frac{z_{11}^{w_2+e}}{n} \right)}{\text{plim} \left[ \left( \frac{z_{11}^*}{n} \right) \left( \frac{z_{11}^{w_2}}{n} \right) - \left( \frac{z_{11}^*}{n} \right) \left( \frac{z_{11}^{w_2+e}}{n} \right) \right]};
\]

\[
\eta_{21} = \frac{\text{plim} \left[ \left( \frac{z_{21}^*}{n} \right) - \left( \frac{z_{21}^{w_2}}{n} \right) \right] \left( \frac{z_{21}^*}{n} \right) \left( \frac{z_{21}^{w_2+e}}{n} \right)}{\text{plim} \left[ \left( \frac{z_{21}^*}{n} \right) \left( \frac{z_{21}^{w_2}}{n} \right) - \left( \frac{z_{21}^*}{n} \right) \left( \frac{z_{21}^{w_2+e}}{n} \right) \right]};
\]

Without loss of generality, we can assume that

\[
\text{plim} \left( \frac{z_{11}}{n} \right) = \text{plim} \left( \frac{z_{12}}{n} \right) = \text{plim} \left( \frac{z_{1e}}{n} \right) = \text{plim} \left( \frac{z_{2e}}{n} \right) = 0;
\]

\[
\text{plim} \left( \frac{z_{22}}{n} \right) > 0
\]

which implies that

\[
\eta_{12} = \frac{\text{plim} \left( \frac{z_{12}^*}{n} \right) \left( \frac{z_{12}^{w_2}}{n} \right)}{\text{plim} \left[ \left( \frac{z_{12}^*}{n} \right) \left( \frac{z_{12}^{w_2}}{n} \right) - \left( \frac{z_{12}^*}{n} \right) \left( \frac{z_{12}^{w_2+e}}{n} \right) \right]};
\]

\[
\eta_{21} = \frac{\text{plim} \left( \frac{z_{22}^*}{n} \right) \left( \frac{z_{22}^{w_2}}{n} \right)}{\text{plim} \left[ \left( \frac{z_{22}^*}{n} \right) \left( \frac{z_{22}^{w_2}}{n} \right) - \left( \frac{z_{22}^*}{n} \right) \left( \frac{z_{22}^{w_2+e}}{n} \right) \right]};
\]

If the denominator is positive (\( z_2 \) close to \( w_2 \) and \( z_1 \) close to \( w_1 \)) and \( \text{plim} \left( \frac{z_{1}(x-w_2)}{n} \right) \) is a better instrument for \( w_2 \) than for \( x \) (note that these assumptions would all be true if \( z_2 = w_2 \) or \( z_2 = x \) and \( z_1 = w_1 \)), then

\[
\eta_{12} < 0, \eta_{21} > 0,
\]

which implies that

\[
\text{plim} \hat{a}_1 < \alpha_1,
\]

\[
\text{plim} \hat{a}_2 > \alpha_2.
\]

In words, the estimate on diagnosis & evaluation would be negatively biased, and the estimate on mental illness or SMI would be biased upwards.

### 8.7 Estimates of the Impact of Covariates

Table A.6 displays the estimates of the effects of demographic characteristics on the propensity to use different services (\( y_{ij}^* \) in equation (1)). For the most part, the observed characteristics do not have statistically significant effects on service receipt, but there are some interesting exceptions. We find that clients with learning disabilities (0.680) and those receiving government assistance (0.491) are more likely to receive diagnosis & evaluation services. The probability of receiving training is higher for persons with government assistance (0.777) but lower for men (−0.315) and for those with musculoskeletal disabilities (−0.489) and/or substance abuse problems (−0.373). The receipt of education increases for those with more
education (0.082) and for those with access to transportation and a driver’s license (0.679). Interestingly, however, there is no statistically significant effect associated with having a serious mental illness or a significant disability.

Table A.7 reports the effects of the demographic, socioeconomic, and disability-related characteristics on the three labor market outcomes of interest ($z_{it}$ in equation (2), $w_{it}$ in equation (3), and $r_{it}$ in equation (4)). For labor market outcomes, almost all of the estimates are statistically significant. Many of the estimates are as expected including positive effects of being white on employment propensity (0.157) and log quarterly earnings (0.362) as well as positive effects of education on employment propensity (0.024) and log quarterly earnings (0.053). The two transportation variables also have positive impacts on both labor market outcomes. The local labor market employment rate increases employment probabilities but decreases conditional earnings, suggesting that it might have been useful to include a measure of local wage rates. Some of the demographic and socioeconomic parameter estimates are counterintuitive. In particular, having a serious mental illness (SMI) increases employment propensity (0.194), receipt of special education services increases log quarterly earnings (0.259), while being married decreases both employment propensity (−0.322) and log quarterly earnings (−0.138). The marriage effects can occur through income effects associated with having a spouse.

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 (−0.210) while increasing log quarterly earnings (0.399). 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, the SMI estimates are explaining a significant amount of variation in labor market outcomes. SMI, by itself, increases employment (0.194) and increases log quarterly earnings (0.964). For males and whites, there are added interaction effects, all adversely affecting labor market outcomes. However, overall, the estimates with respect to SMI effects are hard to explain. Education interacted with SMI has negative effects, and age interacted with SMI has mixed 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. 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

---

**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, and items with + were restricted.
Table A.7: Labor Market and DI/SSI Effects

<table>
<thead>
<tr>
<th>Variable</th>
<th>Employment</th>
<th>Log Quarterly Earnings</th>
<th>DI/SSI Receipt</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Estimate</td>
<td>Std Err</td>
<td>Estimate</td>
</tr>
<tr>
<td>Constant</td>
<td>0.052 **</td>
<td>0.021</td>
<td>5.311 **</td>
</tr>
<tr>
<td>Male</td>
<td>0.024 **</td>
<td>0.006</td>
<td>0.360 **</td>
</tr>
<tr>
<td>White</td>
<td>0.157 **</td>
<td>0.007</td>
<td>0.362 **</td>
</tr>
<tr>
<td>Education</td>
<td>0.024 **</td>
<td>0.001</td>
<td>0.055 **</td>
</tr>
<tr>
<td>Special Education</td>
<td>-0.014 **</td>
<td>0.027</td>
<td>-0.029 **</td>
</tr>
<tr>
<td>Education Missing</td>
<td>0.423 **</td>
<td>0.013</td>
<td>0.626 **</td>
</tr>
<tr>
<td>Age/100</td>
<td>-0.701 **</td>
<td>0.007</td>
<td>0.043 **</td>
</tr>
<tr>
<td>Married</td>
<td>-0.322 **</td>
<td>0.007</td>
<td>-0.138 **</td>
</tr>
<tr>
<td>#Dependants</td>
<td>0.061 **</td>
<td>0.002</td>
<td>0.064 **</td>
</tr>
<tr>
<td>Transportation Available</td>
<td>0.122 **</td>
<td>0.007</td>
<td>0.089 **</td>
</tr>
<tr>
<td>Has Driving License</td>
<td>0.213 **</td>
<td>0.007</td>
<td>0.333 **</td>
</tr>
<tr>
<td>Receives Govt Assistance</td>
<td>-0.332 **</td>
<td>0.008</td>
<td>-0.237 **</td>
</tr>
<tr>
<td>Musculoskeletal Disability</td>
<td>0.097 **</td>
<td>0.007</td>
<td>0.130 **</td>
</tr>
<tr>
<td>Learning Disability</td>
<td>0.406 **</td>
<td>0.012</td>
<td>0.225 **</td>
</tr>
<tr>
<td>Mental Illness</td>
<td>-0.210 **</td>
<td>0.012</td>
<td>0.399 **</td>
</tr>
<tr>
<td>Substance Abuse</td>
<td>0.213 **</td>
<td>0.007</td>
<td>0.086 **</td>
</tr>
<tr>
<td>Disability Significant</td>
<td>-0.027 **</td>
<td>0.009</td>
<td>-0.170 **</td>
</tr>
<tr>
<td>Disability Most Significant</td>
<td>-0.112 **</td>
<td>0.010</td>
<td>-0.280 **</td>
</tr>
<tr>
<td>SMI</td>
<td>0.194 **</td>
<td>0.028</td>
<td>0.964 **</td>
</tr>
<tr>
<td>Male * SMI</td>
<td>-0.094 **</td>
<td>0.012</td>
<td>-0.521 **</td>
</tr>
<tr>
<td>White * SMI</td>
<td>-0.535 **</td>
<td>0.012</td>
<td>-0.482 **</td>
</tr>
<tr>
<td>Education * SMI</td>
<td>-0.019 **</td>
<td>0.001</td>
<td>-0.047 **</td>
</tr>
<tr>
<td>Age/100 * SMI</td>
<td>0.236 **</td>
<td>0.015</td>
<td>-0.083 **</td>
</tr>
<tr>
<td>Local Employment Rate</td>
<td>0.185 **</td>
<td>0.067</td>
<td>-0.185 **</td>
</tr>
</tbody>
</table>

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

For DI/SSI receipt, almost all of the effects are statistically significant and with expected signs, also. For example, the probability a client takes-up DI/SSI is estimated to decrease for white (−0.114), education (−0.044), and transportation available (−0.068). Surprises are married (0.484), learning disability (−0.166), and local employment rate (0.050).

So far all of the discussion has concerned the effect of purchased services on labor market outcomes. In fact, DARS also provides some services in-house, and other services sometimes are paid for by other organizations, and, as discussed in Section 3.1.2, we have some information about those other services. Using this data, we allow the effects of covariates on the receipt of such services to be proportionate to their effect for service choice in equation (1) and their effect for employment propensity in equation (2) as reported in Table 4, for conditional log quarterly earnings in equation (3) as reported in Table 5, and for DI/SSI receipt in equation (4) as reported in Table 6. The estimated proportion for service choice propensity is 0.824** (0.280) which is not significantly different from 1.0. Thus, decisions about using non-purchased services are similar to those for purchased services. By contrast, the estimated proportion for employment propensity, conditional log quarterly earnings, and DI/SSI receipt propensity is 0.432** (0.033) which is significantly different from 1.0. Thus, the effect of non-purchased services on labor market outcomes and DI/SSI receipt is 43.2% of that for purchased services.

8.8 Smoothed Sample DI/SSI Probabilities

Smoothed sample DI/SSI probabilities conditional on predicted probabilities are displayed in Figure 6. As was similar for employment, the predicted probabilities fit pretty well except for values above 0.7.

8.9 Application Test

In this Appendix, we present a test for whether the decision to apply for VR services is exogenous. We begin by formalizing the test, then describe the data, and conclude by summarizing the test results.

8.9.1 Methodology

Let

$$p(x_i, \gamma) = \Phi(x_i \gamma)$$

(13)

be the probability that a disabled person $i$ applies for DARS services where $x_i \in X$ is a set of exogenous variables with at least one variable not directly affecting service choice or labor market outcomes. We have in mind for such an “instrument” the variation in CSB funding on employment services. Divide $X$ into subsets, indexed by $k$. Define $P_k = P(X_k)$ to be the
Figure 6: Smoothed Sample DI/SSI Probabilities Conditional on Predicted Probabilities

proportion of the population belonging to subset $k$, and let $\hat{P}_k$ be a consistent estimate of $P_k$ from some sample. Johnson et al. (2015) provides such an estimate for “composite” CSBs based on work in Stern (2014). Then, the expected number of applicants from subset $k$ is

$$N_k = M \int_{x \in X_k} p(x, \gamma) dF_x(x)$$

where $M$ is the population of disabled people and $F_x(x)$ is the distribution of $x$.

Assume we have a consistent estimate of $M (x) = M f_x (x)$ denoted as $\hat{M}(x) \forall x \in X$. Then we can write the expected number of DARS applicants belonging to subset $k$ as

$$N_k (\gamma) = \int_{x \in X_k} p(x, \gamma) M_x(x) dx$$

and approximate it as

$$\hat{N}_k (\gamma) = \int_{x \in X_k} p(x, \gamma) \hat{M}_x(x) dx.$$

In the data, we observe the actual # of people applying to DARS from subset $k$ which we call $n_k$. Then, a MOM estimate of $\gamma$ is

$$\hat{\gamma} = \min_k \sum_k \omega_k \left[ n_k - \hat{N}_k (\gamma) \right]^2$$

where $\omega_k$ is an appropriately chosen weight.

Let $e_i$ be the error in the latent variable equation corresponding to equation (13). Note that the generalized residual for $e_i \mid \text{apply}$ is

$$\hat{e}_i = E(e_i \mid \text{apply}) = \frac{\int_{e>x, \gamma} e dF_e(e)}{\int_{e>x, \gamma} dF_e(e)} = \begin{cases} \frac{\phi(x, \gamma)}{\Phi(x, \gamma)} & \text{if apply} \\ \frac{-\phi(x, \gamma)}{1-\Phi(x, \gamma)} & \text{if not apply} \end{cases}.$$

Next, let $L_i(u_i)$ be the likelihood function associated with person $i$, conditional on error $u_i$ and application, and let $\hat{u}_i$ be the generalized residual for person $i$. We can test for selection effects into the sample by constructing\footnote{41}$\tilde{\omega}_i$ is probably a vector, so $\tilde{\omega}_{eu}$ is as well. Thus, we need to adjust the equation to include the inverse covariance matrix of $\tilde{u}_i$.}

$$\hat{\rho}_{eu} = \frac{n^{-1} \sum_i \hat{e}_i \hat{u}_i}{\sqrt{n^{-1} \sum_i \hat{e}_i^2 \left( n^{-1} \sum_i \hat{u}_i^2 \right)}}.$$

Under $H_0$, $\text{plim} \hat{\rho}_{eu} = 0$ and

$$\sqrt{n} \left( \hat{\rho}_{eu} - 0 \right) \sim N(0, 1).$$
Table A.8: Weighted Moments of Explanatory Variables from ACS

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mean</th>
<th>Std Dev</th>
<th>Variable</th>
<th>Mean</th>
<th>Std Dev</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female</td>
<td>0.516</td>
<td>0.500</td>
<td>High School Diploma</td>
<td>0.489</td>
<td>0.500</td>
</tr>
<tr>
<td>Age/100</td>
<td>0.084</td>
<td>1.056</td>
<td>College Degree</td>
<td>0.394</td>
<td>0.489</td>
</tr>
<tr>
<td>Black</td>
<td>0.185</td>
<td>0.389</td>
<td>Family Income</td>
<td>84.950</td>
<td>85.216</td>
</tr>
<tr>
<td>Other Race</td>
<td>0.077</td>
<td>0.267</td>
<td>Dummy: Fam Inc &gt; $50K</td>
<td>0.585</td>
<td>0.493</td>
</tr>
<tr>
<td>Hispanic</td>
<td>0.073</td>
<td>0.260</td>
<td>In MSA</td>
<td>0.711</td>
<td>0.453</td>
</tr>
<tr>
<td>Married</td>
<td>0.550</td>
<td>0.497</td>
<td>Health Condition</td>
<td>0.086</td>
<td>0.280</td>
</tr>
<tr>
<td>Divorced</td>
<td>0.108</td>
<td>0.310</td>
<td># ADLs</td>
<td>0.030</td>
<td>0.170</td>
</tr>
<tr>
<td>Widowed</td>
<td>0.058</td>
<td>0.234</td>
<td># IADLs</td>
<td>0.055</td>
<td>0.228</td>
</tr>
<tr>
<td>Veteran</td>
<td>0.116</td>
<td>0.320</td>
<td>Functional Limitation</td>
<td>0.077</td>
<td>0.266</td>
</tr>
</tbody>
</table>

Notes
1) Sample size is 59207.
2) Sample weights are used to replicate the population joint density of variables.
3) Median family income = $62.5K.

Distribution of CSB Ratios

Figure 7: Distribution of CSB Ratios

8.9.2 ACS Data

Data on Virginia residents from the 2012 American Community Survey (ACS) are used to estimate the model described in equation (13) \( N = 64009 \); see Johnson et al. (2015) for details.\(^{\text{42}}\) Moments of the data are presented in Table A.8. The surprising number is the large mean family income of $84.95K. This is caused by a fat right tail; the median family income is only $62.5K.

The state is divided up into 40 public mental health care regions called community service boards (CSBs) (Johnson, et al., 2015). Our estimation strategy relies on the existence of significant variation in the ratios of DARS applicants to people with mental health problems across the CSBs. Figure 7 presents the distribution of ratios. One can see immediately that there is significant variation and that the distribution is continuous and well behaved.

8.9.3 Estimation Results

Table A.9 shows the results of the estimation process associated with equation (14). Each of the estimated coefficients should be interpreted as \( \frac{\partial p^*(x_i, \gamma)}{\partial x_i} \) for each of the variables in the table where \( p^*(x_i, \gamma) \) is the latent variable associated with \( p(x_i, \gamma) \); i.e., \( p(x_i, \gamma) = \Pr[p^*(x_i, \gamma) > 0] \). The estimates in Table A.9 are interesting in their own right in that they provide information about how the propensity to apply for DARS services varies with individual characteristics. Women are more likely to apply than men (0.285), and whites are more likely to apply than blacks (−0.252) or Hispanics (−0.237). Veterans

\(^{\text{42}}\)We use the same 2012 data as Johnson et al. (2015) despite our DARS data coming from 2000 because the earliest ACS data with the necessary variables to predict mental health with any precision is from 2008, and the difference between 2008 and 2012 in the distribution of demographic characteristics was not large.
Table A.9: Estimates of Determinants of Walking in the Front Door

<table>
<thead>
<tr>
<th>Variable</th>
<th>Estimate</th>
<th>Std Error</th>
<th>Variable</th>
<th>Estimate</th>
<th>Std Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>-1.152 **</td>
<td>0.437</td>
<td>High School Diploma</td>
<td>-0.074 **</td>
<td>0.000</td>
</tr>
<tr>
<td>Female</td>
<td>0.285 **</td>
<td>0.000</td>
<td>College Degree</td>
<td>-0.141 **</td>
<td>0.001</td>
</tr>
<tr>
<td>Age/100</td>
<td>-0.146 **</td>
<td>0.000</td>
<td>Family Income/1K</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>(Age/100)**2</td>
<td>-0.089 **</td>
<td>0.000</td>
<td>Dummy: Big Fam Inc</td>
<td>-0.292 **</td>
<td>0.001</td>
</tr>
<tr>
<td>(Age/100)**3</td>
<td>0.069 **</td>
<td>0.000</td>
<td>Health Fair</td>
<td>-0.026 **</td>
<td>0.003</td>
</tr>
<tr>
<td>Black</td>
<td>-0.252 **</td>
<td>0.001</td>
<td>Health Poor</td>
<td>1.445 **</td>
<td>0.005</td>
</tr>
<tr>
<td>Other Race</td>
<td>-0.023 **</td>
<td>0.001</td>
<td>Weight/Height</td>
<td>-0.215</td>
<td>0.178</td>
</tr>
<tr>
<td>Hispanic</td>
<td>-0.237 **</td>
<td>0.002</td>
<td>In MSA</td>
<td>0.090</td>
<td>0.000</td>
</tr>
<tr>
<td>Married</td>
<td>-0.213 **</td>
<td>0.001</td>
<td>Health Condition</td>
<td>0.525 **</td>
<td>0.000</td>
</tr>
<tr>
<td>Divorced</td>
<td>0.126 **</td>
<td>0.001</td>
<td># ADLs</td>
<td>0.148 **</td>
<td>0.000</td>
</tr>
<tr>
<td>Widowed</td>
<td>-0.022 **</td>
<td>0.001</td>
<td># IADLs</td>
<td>0.320 **</td>
<td>0.000</td>
</tr>
<tr>
<td>Veteran</td>
<td>0.107 **</td>
<td>0.001</td>
<td>Functional Limitation</td>
<td>0.488 **</td>
<td>0.001</td>
</tr>
</tbody>
</table>

Notes
1) Age variables are transformed into 1st-, 2nd-, and 3rd-order orthogonal polynomials.
2) Double-starred items are statistically significant at the 5% level.

(0.107) have higher application rates, and application rates decline with education (-0.074, -0.141). Health conditions have large effects on application rates as people in poor health (1.445), with health conditions (0.525), or with functional limitations (0.488) are significantly more likely to apply for DARS services.

The other feature of this estimation process is the extremely high proportion of the data variation explained by the variables in Table A.9. The weighted sum of squared residuals when \( \gamma = 0 \) is 3641.2, and, at \( \hat{\gamma} \) (reported in Table A.9), it is 0.07. This implies that there is very little room for unobserved characteristics, possibly correlated with errors in the model described in the text, to exist. More precisely, there may be other variation \( v_{ki} \) across observations \( i \) with each homogenous cell \( k \) used in equation (14) that average out over the cell and therefore have no effect on estimation. However, if one thinks that the \( N_k(\gamma) \) people from cell \( k \) applying to DARS are those with the greatest (latent) value of participating, then the standard deviation of the aggregated error (across the cell) is the standard deviation of sample average of the of the first \( N_k(\gamma) \) sample order statistics of the cell;

\[
StDev = \sqrt{\frac{1}{N_k(\gamma)} \sum_{(i) \in k} (v_{k(i)} - \bar{v}_k)^2} \tag{15}
\]

where \( (i) \) is the \( v_{ki} \) term for the \( i \)th largest latent value and \( \bar{v}_k \) is the sample average of the \( v_{ki} \)'s associated with the \( N_k(\gamma) \) largest latent values. As an illustration, consider the case where \( v_{ki} \sim iidN(0, 1) \) and \( \gamma = 0 \). Then, the simulated standard deviation in equation (15) is presented in Figure 8. The bars for proportion applying is 1.00 is the standard deviation of the mean: \( 1/M_k \) where

\[
M_k = \int_{x \in X_k} M(x) \, dx
\]

is the number of disabled people in cell \( k \). The bars for other proportions are for different values of \( N_k(\gamma)/M_k \). As seen in Figure 7, the median proportion is approximately 0.06. Thus we simulate results for 0.04, 0.06, and 0.08. In fact, \( \gamma \neq 0 \), causing heterogeneity in the sample and changing the sample of applicants from those with the highest value of \( v_{ki} \) to those with the highest value of the latent variable. The results in Figure 8 show that the standard deviation of the mean value of \( v_{ki} \) among those who apply for DARS services decreases with sample size \( M_k \), decreases with the proportion applying, and increases with heterogeneity. Compared to a random sample of 10000 disabled people (proportion applying = 1.00) with or without heterogeneity, a sample of 10000 with 6% applying and with heterogeneity has a standard deviation of \( v_{ki} \) among those applying 2.5 times larger (0.01 \( \Rightarrow \) 0.025). If the cell sample size is 5000, then the ratio increases to 2.8 (0.014 \( \Rightarrow \) 0.040). Thus, the standard deviation of \( v_{ki} \) consistent with our results based on cell-mean application rates would have to be on the order of 2.8 times smaller than if all disabled people in the cell applied.

\(^{13}\)As can be seen in, e.g., Headrick and Pant (2012), it is quite difficult to analytically evaluate the covariance matrix of multiple order statistics, especially for large sample sizes. Our need for the standard deviation of the mean error conditional on application can be computed using formulae for moments of truncated normal variables available, for example, in Heckman (1979). However, it is completely straightforward and inexpensive (CPU time) to simulate them.

\(^{14}\)Heterogeneity is simulated for this example as \( x = 3U(0, 1) \).
8.9.4 Test Results

Given the estimates in Table A.9, we can construct generalized residuals and use them in the proposed test statistic described in Section 8.9.1. The test statistics and critical values are reported in Table A.10. Note the range of test statistics associated with service choices is (0.006, 0.076), and the range for labor market and SSI/DI outcomes is (0.018, 0.051), all of which are very small. All of the test statistics fall between the 2.5% and 97.5% critical values. Thus, there is no evidence of selection on unobservables caused by the DARS application decision.

References


