The effect of vocational rehabilitation on the employment outcomes of disability insurance beneficiaries: new evidence from Canada

Michele Campolieti[1,2], Morley KL Gunderson[3*] and Jeffrey A Smith[4,5]

* Correspondence:
morley.gunderson@utoronto.ca
[3]Department of Economics and
Centre for Industrial Relations,
University of Toronto, Toronto,
Canada
Full list of author information is
available at the end of the article

Abstract

We estimate the effects of the vocational rehabilitation (VR) program run by the Canada Pension Plan Disability Program using administrative data. Identification relies on "selection on observed variables" plus careful comparison group selection and institutional knowledge regarding sources of conditional variation in participation. We employ several matching and weighting estimators and emphasize flexible conditioning on variables suggested by theory, the institutional setup and the literature. We find modest, and imprecisely estimated, impacts on employment outcomes for men and larger, sometimes statistically significant, impacts for women. A formal sensitivity analysis finds our results are quite robust to lingering selection on unobserved variables.

JEL codes: I38; J08; J24

Keywords: Vocational rehabilitation; Disability; Evaluation; Matching

1. Introduction

Recently, disability policy has placed an increased focus on the employment of persons with disabilities as a way to generate cost savings via reduced benefit payments. Increasing the return-to-work of disability beneficiaries is particularly important because of shifts in the characteristics of persons collecting disability benefits over the last several decades. First, the share of younger persons on the disability rolls is increasing in many countries (Bound and Burkhauser 1999). This demographic change makes it more difficult for disability beneficiaries to use disability insurance as a bridge to retirement benefits. Second, the nature of disabilities has shifted somewhat over the last several decades. Specifically, disability beneficiaries in Canada and the U.S. increasingly suffer from health problems such as musculoskeletal and soft tissue problems as well as mental disorders that tend to be more chronic in nature, rather than health problems associated with higher mortality rates and shorter claim duration (Bound and Burkhauser 1999; Campolieti 2002; Rupp and Scott 1998). These two changes in the nature of the disabled population both increase the time horizon over which investments in the skills of the disabled can pay off and increase the fraction of the disabled likely to see improvements in their medical conditions that would allow for a return to the labor market.

A few policy instruments can be used to increase the attachment of disability beneficiaries to the labor market. On the one hand, disability insurance programs can allow disability beneficiaries to engage in work activities and not terminate their benefits as long as they do not cross an earnings threshold. This sort of policy (referred to as benefit offset in the U.S. and an earnings disregard in Canada) can allow disability beneficiaries to earn some labor market income as well as test the waters to determine whether they can work in a sustained fashion. These sorts of initiatives are often bundled with trial work periods and automatic reinstatement options, where if an individual leaves the disability rolls but then determines they are unable to maintain their employment because of their disability they can resume receiving a disability pension. There is some evidence that such incentives can increase the employment of disability beneficiaries (Campolieti and Riddell 2012; Kostøl and Mogstad 2013) as well as their exit rates from the disability rolls (Kostøl and Mogstad 2013).

On the other hand, the disability insurance program can also use vocational rehabilitation (e.g., training that upgrades existing skills or provides new skills, often in combination with job search assistance) as an alternative policy instrument to increase the employability of disability insurance beneficiaries. While used extensively by workers' compensation programs for many decades (Allingham and Hyatt 1995), these programs are a relatively recent addition to disability insurance programs in many countries. For example, the Canada Pension Plan disability (CPPD) program only introduced a vocational rehabilitation program to facilitate the return-to-work of its beneficiaries in the 1990s, well after the CPPD program was created in the mid-1960s. The U.S. also expanded the potential for vocational rehabilitation in its disability policy via the Ticket to Work (TTW) initiative that provides disability beneficiaries with a voucher that can be used to purchase public or private sector employment services (U.S. Social Security Administration 2004).

Generally, most countries use a combination of these sorts of policies in their disability strategies. Consequently, disentangling the impact of greater vocational rehabilitation services from financial and other incentives to encourage employment can be somewhat difficult because a client may have received vocational rehabilitation services and also have a number of other incentives influencing his or her decision to return to the labor market. Moreover, in some countries, like the United States, there is a complicated structure of programs available to persons with disabilities and disability beneficiaries that makes it difficult to design, test and implement initiatives (Wittenburg et al. 2013).

In contrast, Canada offers a unique opportunity to study the effects of vocational rehabilitation on exit rates from the disability rolls as well as subsequent employment. Canada's disability insurance program, the Canada Pension Plan Disability (CPPD) program, did not introduce its vocational rehabilitation program, the CPPD-VR program, until the mid-1990s. Moreover, when it did introduce this program it did not have any other incentives (e.g., allowable earnings, trial work periods, or automatic reinstatement options for disability beneficiaries who left the disability rolls) available to disability beneficiaries that might increase their employment. However, the CPPD program began to change its strategy and approach for facilitating the attachment of its beneficiaries to the labor market by beginning to introduce these incentives in 2001 (Campolieti and Riddell 2012). Consequently, there is a short window available in the Canadian data where only vocational rehabilitation services were available to disability beneficiaries.

This paper presents estimates of the effect of the vocational rehabilitation program run by the Canada Pension Plan disability (CPPD) program on the labor market outcomes of disability insurance beneficiaries. Our identification strategy relies on "selection on observed variables", bolstered in our case by careful selection of the comparison group and by the institutional knowledge that opportunities for participants in the CPPD VR program do not depend on local labor market conditions. We use administrative data from the CPPD program and obtain our estimates with propensity score matching estimators (kernel and local linear), an inverse probability weighting procedure and a genetic matching estimator. Finally, we also provide some cost-benefit computations from the perspective of the government.

The remainder of the paper unfolds as follows: Section 2 presents background information on the CPPD program and its vocational rehabilitation program. Section 3 provides a detailed description of the data. Section 4 describes our treatment effect estimators. Section 5 presents our principal empirical findings, a sensitivity analysis and a comparison with previous research. Section 6 contains a calculation of the potential savings in transfer payments for the CPPD program resulting from the CPPD-VR program. We conclude the paper with a brief summary of our results and their implications for policymakers.

2. Background information and previous research

The Canada Pension Plan Disability (CPPD) program is the disability component of the Canada Pension Plan (CPP), which was established in 1966. Quebec operates its own program, the Quebec Pension Plan, which also contains a disability component. The CPPD program is quite similar to the U.S. Social Security Administration's Disability Insurance (DI) program. The CPPD program is available to individuals with severe (preventing the individual from working regularly) and prolonged (a long-term condition or condition likely to result in death) disabilities. In addition, applicants to the CPPD program must also satisfy a contribution requirement, which functions as a recency-of-work requirement (all workers in Canada must make contributions). In other words, applicants must have worked in some of the years leading up to an application for disability benefits.

The CPPD program does not place any conditions on eligibility related to the source of the disability. This differs from workers' compensation programs, which focus on the compensation of disabilities arising in the course of employment. Benefits are paid until the maximum age of 65, at which point disability pensions are converted into retirement pensions. The CPPD program reevaluates or reassesses beneficiaries from time-to-time. If this reassessment indicates that an individual no longer has a disability, as defined by the CPPD program, then disability benefits would be terminated as a consequence.

The Canada Pension Plan Disability program Vocational Rehabilitation (CPPD-VR) program was established in 1997 as a successor to the National Vocational Rehabilitation program, which evolved from a pilot program that operated from 1992–1997. The goal of VR under CPPD is to facilitate the client's return to gainful employment and generate cost savings to the CPPD fund. The program is administered by Social Development Canada and was delivered to clients by about 30 case managers in the 1990s. These case managers screen CPPD beneficiaries for their suitability for vocational

rehabilitation and manage the rehabilitation assignments. Specifically, after consulting with the client the case managers develop an individualized plan, which can include a vocational reassessment and planning, skills development and job search assistance. Third-party contractors provide the services and in some cases may also have input into the design of the client's vocational rehabilitation plan.

While the CPPD-VR program is relatively new, VR has been used by other disability programs much more extensively. For example, workers' compensation programs (especially in Canada) often have a strong emphasis on VR in their rehabilitation strategies (Allingham and Hyatt 1995). A number of studies have looked at the effects of VR provided by workers' compensation programs (primarily) and other programs on the labor market outcomes of individuals using data from the United States as well as Canada (e.g., among others, Dean and Dolan 1991a, b; U.S. General Accounting Office 1987, 1994; Gardner 1988; Skaburskis and Collignon 1991; Allingham and Hyatt 1995; Campolieti and Hyatt 2011; Dean et al. 2013a, b). This previous literature has produced a large range of estimates, e.g., some studies indicate that VR is quite effective in improving labor market outcomes, while others find that VR had little or no effect on the labor market outcomes of program participants. However, the interpretation of these findings is complicated by the differences in programs, which may also have different mandates and strategies for rehabilitation, across studies as well as the subject populations (e.g., disability insurance or worker's compensation recipients versus persons with disabilities who do not collect disability benefits). Another potential concern in many of these earlier studies is that there may be identification problems since many were primarily descriptive (e.g., U.S. General Accounting Office 1987, 1994; Gardner 1988; Skaburskis and Collignon 1991) or utilized bivariate normal selection models or selection models using other distributional assumptions with no (or questionable) exclusion restrictions (Allingham and Hyatt 1995; Campolieti and Hyatt 2011). These identification problems may create biased estimates of the impact of VR on employment outcomes that make it more difficult to make conclusive statements about the effectiveness of VR initiatives, especially when combined with differences in programs and study populations.

Wittenburg et al. (2013) review the results from a number of evaluations of demonstration and employment programs for disability beneficiaries and persons with disabilities in the United States, which include the Ticket to Work as well as earlier programs that provided vocational rehabilitation and other supports to various populations of persons with disabilities. They focus on studies using (in their view) rigorous evaluation methods (i.e., primarily relatively compelling non-experimental research designs along with some experiments) so as to minimize identification problems in the interpretation of estimates. At the same time, they consider a broad range of studies and interventions, some of which involve only counseling and case management while others include training. In addition, the programs they consider serve a broad range of study populations and groups. Overall, they did not find that there was much evidence of an effect of these interventions on program participants. However, their review also suggests that interventions targeted at some subpopulations, e.g., younger persons and persons with mental health impairments, are more successful. However, as we noted earlier, they also highlight the difficulties associated with evaluating specific programs in the context of a highly complicated institutional thicket involving multiple agencies and numerous programs and policies.

Aakvik et al. (2005) is one of the few non-experimental studies in the existing literature on VR programs that addresses many of the problems in the earlier studies (see also the more structural Dean et al. (2013a, b) papers). Their identification strategy builds on the assumption that the availability of training slots in the VR program is not correlated with economic conditions in a program participant's home district. Aakvik et al. (2005) study the VR program in Norway, which pays income support as well as providing training programs to individuals who are unable to return to work after 52 weeks on sickness benefits. Their findings suggest that women who received VR have higher employment rates than those who did not receive training, but their estimates are not very precise. They also conclude that the gains in employment for program participants would be increased if the least employable were encouraged to participate in the program rather than using the existing selection rule that encouraged persons who are deemed to have good employment prospects, i.e., cream-skimming, to participate in the program.

3. Data

The data used in this analysis come from the administrative records of the Canada Pension Plan Disability program. The data were drawn from the following files: 1) Master Benefits File (MBF), which includes information on CPPD beneficiaries, such as individual characteristics; 2) Rules Based Reassessment System (RBRS) or reassessment file, which includes information on beneficiaries identified for reassessment, beneficiaries who have been reassessed and those who reported a return to work; 3) Rehabilitation Case Management System (RCMS), which includes a wide range of information on each individual (demographics, type of disability, etc.) who was enrolled in the CPPD-VR program; and, 4) Record of Employment Master File (ROEMF), which contains annual information on total income, labor market earnings and contributions to the CPP program from 1990 until 2001. We excluded people who died or turned 65 during the years we focus on since they are no longer eligible for CPPD.

The treatment group in our analysis is defined as the cohort of individuals who started the CPPD-VR program during 1998. Some of the individuals in this cohort dropped out of the program and so failed to complete their VR assignments. As we define treatment as starting VR, these individuals remain in our treatment group. We track the post-VR experiences for the treatment group up until 2001. As we discussed earlier, the CPPD program began introducing greater incentives to encourage return to work during and after 2001. Campolieti and Riddell (2012) found that these initiatives were associated with an increase in the employment of disability beneficiaries. Consequently, including data after 2001 would make it difficult to distinguish the effect of VR from these new initiatives. However, we do have a window, i.e., 1998 to 2001, to estimate impacts of the CPPD-VR program that are not contaminated by any other policy. Prior to 1997, VR was offered through the National Vocational Rehabilitation program. The CPPD-VR replaced this program in 1997, but was still ramping up from the previous program. We focus on the 1998 cohort because potential VR clients and case managers would have become more accustomed to the new program structure by 1998, relative to clients and case managers in 1997.

As noted in Dean et al. (2013a, b), one could make a distinction between the short- (up to eight quarters after treatment) and long-run (more than eight quarters after

treatment) effects of VR. The short window for our outcome measures (up to three years) means that we consider only the short-run effects of the CPPD-VR program. Viewed as proxies for the long-run estimates that we would prefer to have, our estimates may embody a downward bias if the VR treatment itself takes a long time, and/ or if treated individuals who return to employment only slowly find their footing in the labor market. We can rule out the first of these concerns because the VR program we consider provides services in a concentrated, contiguous block, rather than spreading them out over time as in Dean et al. (2013a, b).

We draw our comparison group from the reassessment file (RBRS) and include individuals who entered the reassessment file in 1998. Individuals are reassessed in the CPPD if they are believed to have a high probability of returning to work. Individuals in the reassessment file fit into two broad classifications. The first type of reassessment includes those who had been flagged during their initial application for CPPD benefits as being likely to regain their earnings capacity and be able to return to work (i.e., those initially flagged for reassessment). The second type of reassessment includes persons who are reassessed because of their earnings (i.e., reassessed because of earnings). The CPPD program monitors its beneficiaries for earnings and for contributions to Canada's unemployment insurance program (called "Employment Insurance", or EI) through information sharing agreements with the EI program and with Revenue Canada (the government tax agency). Those who are deemed to have earnings or made EI contributions, based on information in these administrative sources, are reassessed. About 46 percent of the individuals in our comparison group got flagged for reassessment at the time of initial application while the remainder were reassessed due to their earnings. We focus on 1998 in order to temporally align the treatment and comparison groups. Temporal alignment ensures that treated and untreated individuals face the same general economic conditions and similar program environments. As individuals in the RBRS database have either returned to work, been reassessed for their potential to return to work or are being reassessed for their potential to return to work, we view them as more comparable to the VR treatment group than other CPPD beneficiaries. We think this substantially reduces the selection problem facing our "selection on observed variables" identification strategy; to the extent that selection bias remains after our conditioning, we expect our comparison group to look "too good" and so bias our estimated treatment effects downward. Like the treatment group, we only track our comparison group through 2001.

Dean and Dolan (1991a, b) argue that the preferred comparison group in an evaluation of a VR program consists of clients who enroll in VR services but drop out prior to completion. They argue that using dropouts reduces concerns about selection bias because dropouts and completers share the motivation to apply for the VR program, satisfy the relevant eligibility criteria, and likely have similar levels of (unobserved) severity in their disabilities. We understand their reasoning, but note three concerns with this strategy in our context, one practical and two conceptual. At a practical level, we simply do not have enough dropouts to have sufficient statistical power to detect effects of reasonable size. More conceptually, using dropouts changes the nature of the estimand from the effect of starting VR to the effect of finishing VR. These differ to the extent that partial receipt of VR affects outcomes. In addition, using dropouts may accentuate some selection problems (why do the dropouts drop out?) while at the same

time leaving aside potentially desirable comparison group members who do not attempt VR for reasons unrelated to their potential outcomes.

The administrative records contain extensive information on CPP disability insurance recipients. This includes: the age at the onset of the disability, gender, pre-disability educational attainment, province of residence, earnings before entering the disability rolls, earnings after leaving the disability rolls (if they exited), and principal health problem (derived from ICD9 codes). For persons enrolled in VR we also have the total cost of the services provided to these individuals.

We examine the effect of VR on several outcome measures. First, we consider the effect of VR on exit from the disability rolls during the period covered by our data. This indicator takes the value one if the individual left the disability rolls (i.e., stopped receiving CPP disability benefits). While we would worry about a "leaving the rolls" outcome in the context of other programs with frequent turnover, such as social assistance in Canada or food stamps in the US, individuals who exit the CPPD rolls rarely return, particularly in the short run.

Second, we also consider employment-related outcomes based on two definitions of employment. The first simply captures any employment at any time during the period covered by our data, with no restrictions on the level of earnings. The second measure is defined as substantial gainful employment (SGO), where the individual must have earnings above a certain threshold in one of the calendar years covered by our data[1]. In both cases we code employment based on information in the ROEMF, which contains total calendar year labor market earnings as reported on the T4 tax form[2]. We use these definitions to define indicator variables corresponding to our two measures of employment.

4. Methodology

4.1 Estimating the treatment effect

Following the standard notation, let Y_1 denote the outcome for someone who receives the treatment and Y_0 the outcome for someone who does not receive the treatment. Y_1 and Y_0 are potential outcomes because we only observe one for each individual. Let $T = 1$ indicate whether the person received the treatment, i.e., they received some VR, while $T = 0$ denotes that the individual is in the comparison group. The average treatment effect on the treated (ATET) in the population equals $E[Y_1 - Y_0 \mid T = 1] = E[Y_1 \mid T = 1] - E[Y_0 \mid T = 1]$. We estimate this treatment effect using two propensity score matching estimators, a reweighting estimator based on the estimated propensity score and a genetic matching estimator.

We adopt a "selection on observed variables" strategy (Heckman and Robb 1985) to identify the ATET. This approach requires the conditional independence assumption (CIA), i.e., conditional on a set of observed covariates the untreated outcome is independent of participation in the VR program, which is expressed mathematically as $Y_0 \perp T \mid X$. Note that we need assume conditional independence only for the untreated outcome due to our interest in the ATET; as such, we need not rule out selection into treatment based on the treated outcome. The propensity score is the conditional probability of participating or receiving the treatment, i.e., $P(X) = \Pr[T = 1 \mid X]$. We also require overlap or common support in $P(X)$. We assume this holds in the population and then restrict the analysis so it holds in the sample as well. Matching and reweighting

estimators require that control variables (X) used in the analysis satisfy these conditions. We use economic theory, institutional considerations and a review of previous empirical research to determine the variables or factors that affect participation in VR and the labor market outcomes we consider for our study population, i.e., persons with disabilities who are collecting disability benefits. Our discussion provides a justification for the variables we include in our specification of the propensity score. In some cases, our data do not contain direct measures of these factors; in these cases we attempt to justify the proxies we use in their stead.

Following e.g. Smith and Todd (2005), estimating the ATET requires an estimate of the counterfactual E $[Y_0 \mid T = 1]$. Matching estimators estimate the counterfactual using some estimator of the form

$$\hat{E}\left[Y_0 \mid \hat{P}(X_i)\right] = \sum_{j=1}^{J} w\left(\hat{P}(X_i), \hat{P}(X_j)\right) Y_{0j}, \tag{1}$$

where $\hat{P}(X_i)$ is the estimated propensity score for individual i in the treated group, $\hat{P}(X_j)$ is the estimated propensity score for individual j in the comparison group, j indexes all the individuals in the comparison group, and $w(\bullet, \bullet)$ is a weighting function based on the distance between propensity scores for individual i and the individuals in the comparison group. The form of the weights depends on (one might even say defines) the particular matching estimator used.

We use three matching estimators - kernel matching, local linear matching and genetic matching, along with Inverse Probability Weighting (IPW). Using multiple estimators with different strengths and weaknesses but all based on the same identifying assumption of "selection on observed variables" provides a clear indication of the sensitivity of the empirical findings to details of the estimation strategy.

Local linear and kernel matching estimators are special cases of the general class of estimators that uses local polynomial regression to estimate the (conditional) expected value of the counterfactual outcome (Heckman et al. 1997). Both the kernel and local linear matching methods take locally weighted averages of the observations in the non-treated group to construct the counterfactual. For example, the kernel estimator computes the weights in equation (1) as

$$w\left(\hat{P}(X_i), \hat{P}(X_j)\right) = \frac{G\left(\frac{\hat{P}(X_i) - \hat{P}(X_j)}{a_N}\right)}{\sum_k G\left(\frac{\hat{P}(X_k) - \hat{P}(X_j)}{a_N}\right)} \tag{2}$$

where $G(\bullet)$ is a kernel function, $\hat{P}(\bullet)$ is the estimated propensity score, and a_N is a bandwidth. The literature suggests that the choice of kernel does not matter much in practice; we use the Epanechnikov kernel. In contrast, the literature clearly indicates that bandwidth selection does matter; we selected our bandwidths using leave-one-out cross-validation as in Black and Smith (2004). We present the bandwidths we use in the kernel and local linear matching approaches in Table 1.

As noted in e.g. Todd (2008), local linear matching has some advantages over standard kernel matching. In particular, the local linear approach has a faster rate of convergence near boundary points, which is important when the data feature many propensity scores near zero or one (which is not the case in our application). More

Table 1 Bandwidths used for kernel and local linear matching estimators

	Men		Women	
	Kernel	**Local linear**	**Kernel**	**Local linear**
	Sample restricted to propensity score values [0.001, 0.40]		Sample restricted to propensity score values [0.001, 0.35]	
Leaving disability rolls	0.03	0.04	0.02	0.07
Gainful employment	0.03	0.04	0.02	0.08
Substantial gainful employment	0.02	0.04	0.01	0.07
	Sample restricted to propensity score values [0.001, 0.35]		Sample restricted to propensity score values [0.001, 0.30]	
Leaving disability rolls	0.03	0.04	0.01	0.06
Gainful employment	0.02	0.04	0.01	0.07
Substantial gainful employment	0.01	0.04	0.01	0.07
	Sample restricted to propensity score values [0.001,0.50]		Sample restricted to propensity score values [0.001, 0.40]	
Leaving disability rolls	0.02	0.04	0.02	0.07
Gainful employment	0.02	0.04	0.01	0.08
Substantial gainful employment	0.02	0.04	0.01	0.07

Note: We use an Epanechnikov kernel and select the bandwidth using leave-one-out cross validation.

generally, and more relevant for us, estimating a local slope parameter yields better estimates when the untreated outcome varies strongly with the estimated propensity score (and when, as one would expect, the treatment and control observations have different propensity score distributions). The cost of the local linear estimator is the degree of freedom lost to estimating a slope coefficient in addition to an intercept in each local regression. We used the STATA ado file (psmatch2) developed by Leuven and Sianesi (2003) for the kernel and local linear matching estimators.

As discussed in e.g. Busso et al. (2013), IPW reweights the comparison observations using the estimated propensity score to implicitly create a comparison sample with the same distribution of observed characteristics as the treatment group. IPW has two important advantages relative to the other estimators we employ: First, it just reweights the data using the (estimated) propensity score and so does not require the (often tiresome and sometimes problematic) choice of a bandwidth. Second, under certain assumptions, the inverse probability weighting estimator achieves the semi-parametric efficiency bound derived by Hahn (1998).

We check for covariate balance using the approach in Dehejia and Wahba (1999) as implemented in Becker and Ichino (2002); see Lee (2013) for a recent overview on balancing tests. This process involves manually specifying and checking the propensity score to ensure balance. While our final specification achieves balance, getting to that specification required a series of modifications to the model until balance was achieved (i.e., adding some higher order polynomials and interaction terms). We used the estimated propensity scores from the resulting model for the kernel and local linear matching estimators and for IPW.

The Genetic Matching (Genmatch) algorithm of Diamond and Sekhon (2013) provides an alternative path to covariate balance. This approach uses an evolutionary search algorithm (i.e. a genetic algorithm) that chooses a weighting matrix W that maximizes the balance of the observed covariates across the treatment and comparison

groups in the context of single nearest neighbor matching with replacement, by minimizing the following criterion.

$$\sqrt{\left(X_i - X_j\right)'\left(S^{-\frac{1}{2}}\right)' W S^{-\frac{1}{2}}(X_i - X_j)}. \qquad (3)$$

In (3), S is the covariance matrix of X, and $S^{-\frac{1}{2}}$ is the Cholesky decomposition of S. We follow Diamond and Sekhon (2013) and include in X the variables in our final propensity score model along with the estimated scores. We obtain our genetic matching estimates and related statistics using the R package *Matching* developed by Sekhon (2011).

We present separate estimates for men and women because there can be differences in the labor market attachment of males and females as well as the type of work they do. These differences are confirmed by a likelihood ratio test which rejects the null hypothesis that the parameters of the propensity score model for males and females are the same (likelihood ratio test statistic of 52 with a p-value of 0.003). The standard errors on the ATET for our kernel matching, local linear matching and inverse probability weighting estimators are all computed by bootstrapping.

4.2 The propensity score specification

Our specification of the propensity score includes a variety of different variables. We group these variables as individual characteristics, health information, institutional information and economic variables. Our rationale for including these variables in the propensity score model draws on economic theory, previous empirical research and institutional knowledge.

Our controls for individual characteristics include age at the onset of the disability, marital status at time of application, an indicator for children, and educational attainment. Public health researchers have found a strong association between age and disability and the recovery from functional limitations caused by disabilities (e.g. Beckett et al. 1996; Anderson et al. 1998) and this could influence the degree to which individuals participate in VR as well as their likelihood of employment. Marital status and the presence of children likely also affect participation in VR as well as employment outcomes (Allingham and Hyatt 1995; Aakvik et al. 2005). On average, we expect individuals with higher levels of education to have higher opportunity costs associated with remaining on CPPD due to more favorable labor market opportunities as well as lower psychic costs of participation (e.g., Allingham and Hyatt 1995).

An individual's health problem is also an important control variable in the propensity score model for VR participation. Different health problems imply varying degrees of physical and mental limitations, which we would expect to affect the degree to which individuals take up VR (Aakvik 2001), their probability of employment (Currie and Madrian 1999) and the type of work they do (and thus their earnings) conditional on employment (Daly and Bound 1996; Campolieti 2009).

Other conditioning variables include the earnings before an individual began collecting disability benefits and the duration of their CPPD spell as of the start of our data. We expect both prior earnings and time on disability to proxy for otherwise unobserved variables such as motivation, determination or attractiveness. For example, if

highly motivated and determined workers do persistently well in the labor market, conditioning on earlier labor market outcomes should help to remove selection bias that results from motivation or determination playing a role in receipt of VR. Indeed, the literature suggests that prior outcomes often play a critical role in the plausibility of a "selection on observed variables" identification strategy. For this reason, we condition flexibly on pre-program earnings. In particular, we use a cubic polynomial of earnings 1-year prior to applying for CPPD benefits, a cubic polynomial of earnings 2-years prior to applying for CPPD benefits, interactions between earnings 1- and 2-years prior to applying for CPPD benefits as well as indicators for zero earnings 1- and 2-years prior to applying for CPPD benefits. This sort of conditioning set is in the spirit of what has been done in the literature evaluating the Workforce Investment Act using similar methods by, e.g., Heinrich et al. (2013) and Andersson et al. (2013). Aakvik et al. (2005) also include some pre-program variables in the conditioning set in their analysis of the Norwegian VR program as do Dean et al. (2013a, b) in their study of VR in Virginia.

Dolton and Smith (2011) study the New Deal for Lone Parents program in the U.K. and find that the durations of the social assistance spells of the program participants and non-participants contain a great deal of information on variables that would otherwise be unobserved. In our study, spell duration on the CPPD (i.e., the length of time the person has received disability benefits as of the start of our data) proxies for the severity of the disability, which we do not observe in our data. Severity has been found to be an important factor affecting the employment of persons with disabilities (Meyer and Mok 2013) as well as the uptake of VR (Allingham and Hyatt 1995). In addition, CPPD spell duration can also proxy for the extent of depreciation in the claimant's human capital stock, which we do not otherwise observe in our data, and which we also expect to have an effect on labor market outcomes and participation in VR (Grossman 1972; Campolieti and Hyatt 2011).

Finally, the provincial unemployment rate captures the economic conditions facing CPPD recipients seeking to return to work. We measure the unemployment rate at the time the individual enrolled in the VR program in the treatment group and the time the individual was flagged for reassessment in the comparison group. The literature on active labor market programs emphasizes the importance of including a measure of local labor market conditions in the propensity score (e.g., Heckman et al. 1997).

We estimate the conditional probability of participation, i.e., the propensity score, with a logit functional form. Generally, the literature has found that matching estimates are not sensitive to the choice of a logit or probit functional form. However, Dolton and Smith (2011), like many other researchers, find that allowing for some flexibility in how the conditioning variables enter the propensity score can be quite important. Consequently, we enter some of variables in our model as cubic polynomials (age at onset of the disability, earnings 1- and 2-years before application, CPPD spell duration and unemployment rates) and also include some interaction terms (an interaction between 1-year and 2-years before application earnings).

Another issue of concern with matching and reweighting estimators of treatment effects is the degree of overlap in the distribution of the propensity score. The literature offers a variety of different ways to measure the region of common support and to impose a common support condition on the analysis. Crump et al. (2009) show that trimming the propensity score can reduce the asymptotic variance of the estimated

Figure 1 Histogram of propensity score for comparison and treatment groups, men.
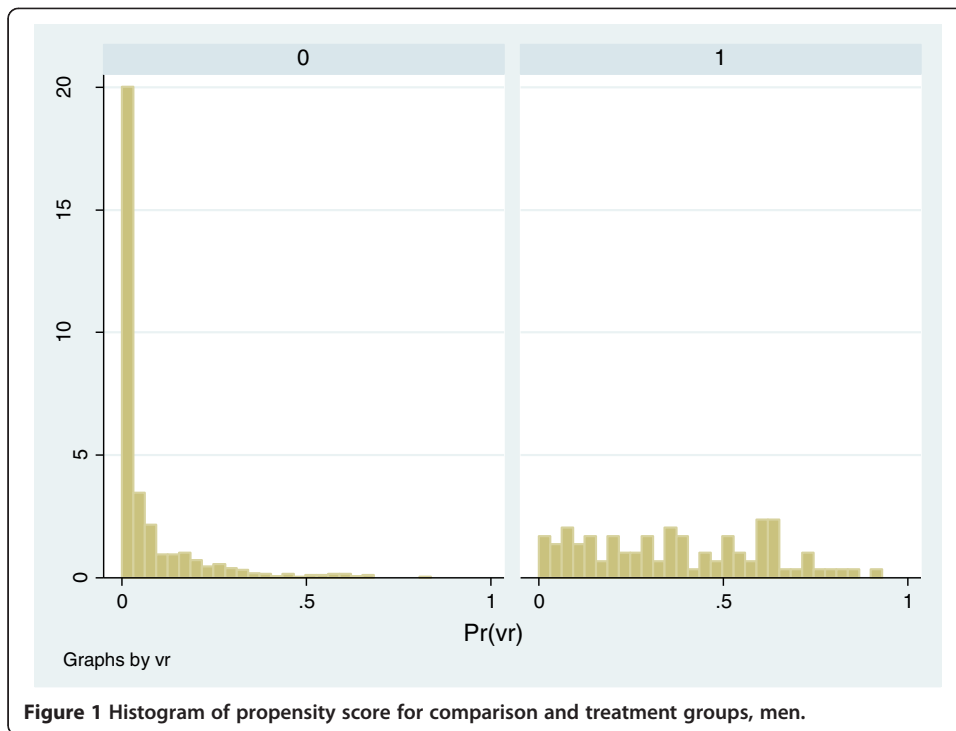
treatment effect, while at the same time, of course, changing its interpretation. More-over, values of the propensity score near zero or one can lead to instability in the esti-mates as well as poor finite sample performance of the IPW estimator (Busso et al. 2013). We present plots of the propensity score by treatment and comparison group for men and women in Figures 1 and 2. These figures provide reason for concern about the degree of overlap in the upper tails of the estimated propensity score distribution for both males and females. Consequently, we trim the data in our analyses in order to

Figure 2 Histogram of propensity score for comparison and treatment groups, women.
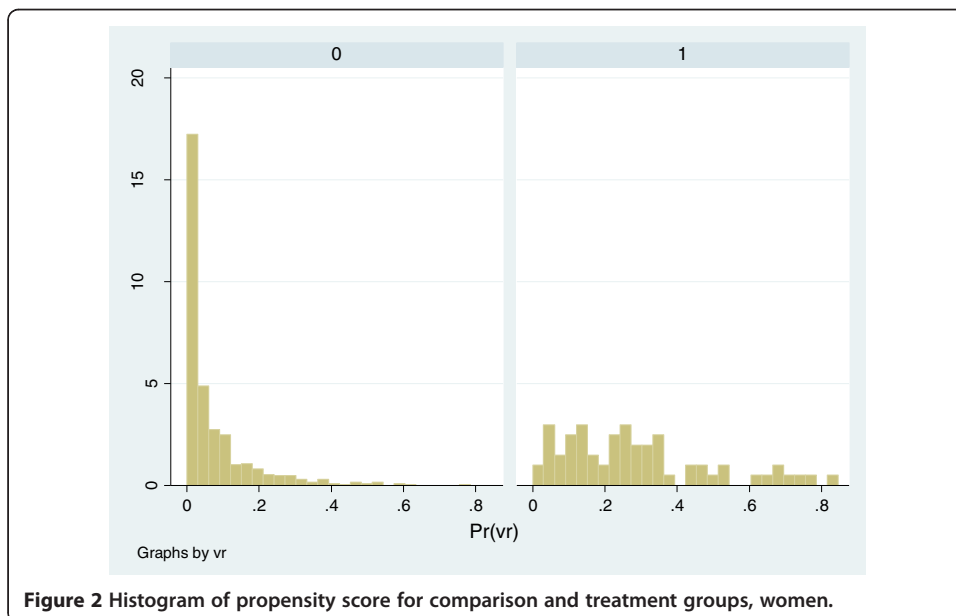
achieve a common support and to avoid issues with instability due to very small estimated propensity scores. Our preferred estimates for men restrict the propensity score to lie in [0.001, 0.40], while for women they restrict the propensity score to lie in [0.001, 0.35]. As a sensitivity analysis, we also consider estimates with less aggressive trimming, so that the propensity scores lie in [0.001, 0.50] for men and [0.001, 0.40] for women, and more aggressive trimming, so that the propensity scores lie in [0.001, 0.35] for men and [0.001, 0.30] for women.

4.3 Program framework

We now present some details on the selection of individuals into the program in order to provide context for the estimates and to help justify the "selection on observed variables" assumption underlying the estimators we use. A recent survey of case managers indicated that many of the individuals in the CPPD-VR program contacted SDC and enrolled themselves (Social Development Canada 2004). Other individuals in the VR program may have enrolled in the program after being referred to the program during a reassessment of their case by CPPD personnel. Quite importantly, the CPPD program differs from some workers' compensation programs, which can compel an individual to enroll in a VR program, because enrollment in the program is voluntary. The CPPD-VR program's requirements in terms of potential clients are that individuals are motivated to return to work and that the client's physician agrees that he or she can cope with a work-related VR program. This suggests that while many individuals may desire to enter the VR program, medical assessments will partly determine who enrolls in the program and who does not. We think this helps with our selection problem by making the VR participants look more like our comparison group members who get reassessed because of the nature of their disability or their demonstrated ability to generate earnings.

Between 1998 and 2003 the average annual expenditures of the CPPD-VR program were about \$4.2 million, so the overall size of the program is small relative to the total expenditures of the CPPD program of several billion dollars (Subcommittee on the Status of Persons with Disabilities 2004). Consequently, there are only a limited number of spots in the CPPD-VR program. The program is not an entitlement, and the number of available slots varies across geographic regions in ways not systematically related to local economic conditions. Moreover, access may depend on caseworker approval; CPPD recipients get assigned to caseworkers in ways unrelated to their likelihood of employment. The allocation of VR services to CPPD recipients thus differs importantly from the way in which Canada allocates training in its active labor market programs. In those programs, opportunities for training do depend on local economic conditions, in addition to caseworker approval. In our context, both the non-systematic assignment of clients to caseworkers and the fact that the number of VR slots in a region do not depend on local economic conditions generate useful (i.e. not correlated with the untreated outcome) variation in treatment status conditional on observed characteristics and so increase the plausibility of our identification strategy.

These arrangements for the CPPD VR program resemble those in Norway, whose VR program is studied by Aakvik (2001) and Aakvik et al. (2005). In Norway, participant slots in general training programs are related to economic conditions, but not in the disability insurance system's VR program. The availability of VR opportunities in the

Norwegian program is related to the capacity of the program in a region. These arrangements in Norway were key to the identification strategy used by Aakvik et al. (2005), which relied on medical reasons determining the enrollment into the VR program and the availability of opportunities in the VR program to be related to the capacity of program in a region (and not the unemployment rate in that region).

5. Empirical results

5.1 Characteristics of the sample and estimates of propensity score

Tables 2 and 3 provide the descriptive statistics for our data for males and females disaggregated by treatment status. The mean age of participants in the VR program is about seven years lower than those in the comparison group (mid-30s for the treatment group and early-40s for the comparison group) for both males and females. For males the proportions of the treatment and comparison groups that are married is almost identical. However, for women only 51 percent of the treatment group is married, while 63 percent of the comparison group is married. The treatment group also has a higher

Table 2 Descriptive statistics, men

	Treatment		Comparison	
	Mean	Standard deviation	Mean	Standard deviation
Age (at onset of disability)	35.726	7.094	43.986	9.409
Married	0.632	0.485	0.657	0.475
Have children	0.632	0.485	0.450	0.498
[Less than high school]				
High school	0.663	0.475	0.206	0.405
Post-secondary	0.105	0.309	0.075	0.263
University degree	0.084	0.279	0.074	0.261
[Other]				
Infectious and parasitic diseases	0.042	0.202	0.037	0.190
Cancer	0.032	0.176	0.076	0.265
Blood diseases	0.000	0.000	0.000	0.000
Mental disorders	0.116	0.322	0.258	0.438
Diseases of the nervous system	0.274	0.448	0.087	0.282
Circulatory diseases	0.084	0.279	0.155	0.362
Respiratory diseases	0.011	0.103	0.016	0.125
Diseases of digestive system	0.032	0.176	0.026	0.159
Genitourinary system diseases	0.011	0.103	0.033	0.178
Musculoskeletal and soft-tissue disorders	0.179	0.385	0.187	0.390
Congenital diseases	0.011	0.103	0.008	0.089
Zero earnings 1-year prior to application	0.347	0.479	0.293	0.455
Zero earnings 2-years prior to application	0.337	0.475	0.336	0.473
Average earnings 1-year prior to application (dollars)	10421.88	13710.64	16132.02	19302.91
Average earnings 2-years prior to application (dollars)	13089.77	15690.21	17829.55	20587.37
Provincial unemployment rate	8.058	1.802	8.316	2.699
Duration on CPPD (days)	1131.295	586.046	1149.897	617.988
Number of observations	95		884	

Table 3 Descriptive statistics, women

	Treatment		Comparison	
	Mean	**Standard deviation**	**Mean**	**Standard deviation**
Age (at onset of disability)	36.552	7.361	43.423	8.668
Married	0.507	0.504	0.618	0.486
Have children	0.522	0.503	0.410	0.492
[Less than high school]				
High school	0.552	0.501	0.325	0.469
Post-secondary	0.164	0.373	0.143	0.350
University degree	0.149	0.359	0.097	0.296
[Other]				
Infectious and parasitic diseases	–	–	–	–
Cancer	0.030	0.171	0.082	0.275
Blood diseases	0.000	0.000	0.000	0.000
Mental disorders	0.254	0.438	0.390	0.488
Diseases of the nervous system	0.045	0.208	0.107	0.309
Circulatory diseases	0.090	0.288	0.052	0.222
Respiratory diseases	–	–	–	–
Diseases of digestive system	0.090	0.288	0.017	0.131
Genitourinary system diseases	0.030	0.171	0.010	0.100
Musculoskeletal and soft-tissue disorders	0.299	0.461	0.263	0.441
Congenital diseases	0.015	0.122	0.013	0.113
Zero earnings 1-year prior to application	0.313	0.467	0.251	0.434
Zero earnings 2-years prior to application	0.254	0.438	0.228	0.420
Average earnings 1-year prior to application (dollars)	10796.25	13260.08	13778.91	15603.86
Average earnings 2-years prior to application (dollars)	13730.81	14902.95	16427.55	16576.09
Provincial unemployment rate	8.012	2.604	8.101	2.350
Duration on CPPD (days)	1242.254	587.261	1024.588	551.253
Number of observations	67		692	

level of educational attainment and a higher proportion of persons with children for both males and females.

The primary medical problems differ somewhat between the treatment and comparison groups as well as by gender. For example, in Table 2, males with diseases of the nervous system comprise 26 percent of the treatment group, but nine percent of the comparison group; mental disorders are 11 percent of the treatment group, but 25 percent of the comparison group; and musculoskeletal and soft-tissue disorders are about 21 percent of both the treatment and comparison groups. In Table 3, the summary statistics for women indicate some different patterns: diseases of the nervous system are four percent of the treatment group, but 10 percent of the comparison group; mental disorders are 26 percent of the treatment group, but 38 percent of the comparison group; and musculoskeltal and soft-tissue problems are 30 percent of the treatment group and 27 percent of the comparison group.

The average earnings before entering the CPPD program are lower in the treatment group than the comparison group. Moreover, the differences in average earnings between the treatment and comparison groups are smaller for women. Regional

unemployment rates are relatively similar between the treatment and comparison groups. The average time on the CPPD program is also similar for men (1276 days for the treatment group, and 1260 days in the comparison group) and somewhat similar for women (1282 days for the treatment group and 1154 days for the comparison group).

We present the standardized differences between the treatment and comparison groups before and after matching in Tables 4 and 5. We present these estimates based on the IPW and the genetic matching algorithm for both men and women. Lee (2013) notes that there is no consensus on how to best show balance and that there are a number of ways to test for balance. However, standardized differences are straightforward to implement and relatively standard in the literature, so we use them. Not surprisingly, the genetic matching algorithm tends to improve covariate balance relative to IPW for many of the observed covariates.

Table 6 presents coefficient estimates from the propensity score models for males and females. For males, higher levels of educational attainment are associated with increases in the probability of participating in the CPPD-VR program. A few of the controls for health problems are associated with decreases in the probability of participating in CPPD-VR, relative to the omitted health problems category (endocrine

Table 4 Standardized differences in treatment and comparison group, inverse probability weighting

	Men		Women	
	Before matching	After matching	Before matching	After matching
Age	−67.9	12.7	−61.2	1.8
High school	66.3	0.7	35.7	4.1
College	13.6	3.8	5.5	−4.8
University	3.3	−6.9	13.7	−1.1
Married	−4.6	6.1	−8.1	−7.8
Have children	6.7	−4.6	20.8	3.5
Infectious and parasitic diseases	8.1	−1.6	–	–
Cancer	8.2	−2.5	−14.0	−1.8
Mental disorders	−14.6	−0.7	−17.8	−6.8
Diseases of the nervous system	16.9	−2.5	−20.8	−2.0
Circulatory diseases	−18.8	8.8	−1.9	−4.8
Respiratory diseases	0.0	−7.6	–	–
Diseases of digestive system	−5.3	−13.4	20.3	23.5
Genitourinary system diseases	−9.3	−1.3	5.9	−6.0
Musculoskeletal and soft-tissue disorders	−4.4	−1.5	14.2	6.2
Congenital diseases	9.5	14.1	4.7	2.9
Zero earnings 1-year prior to application	9.5	6.2	21.8	10.8
Zero earnings 2-years prior to application	11.6	21.6	14.5	2.4
Earnings 1-year prior to application	−18.7	1.0	−9.5	−0.4
Earnings 2-years prior to application	−22.7	−13.9	−3.6	5.8
Unemployment rate	0.2	−0.2	−19.0	−8.2
Duration on CPPD program	−9.7	1.7	18.2	−2.2

Notes: Entries in the table are standardized differences between the treatment and comparison groups.

Table 5 Standardized differences in treatment and comparison group, genetic matching

	Men		Women	
	Before matching	**After matching**	**Before matching**	**After matching**
Age	−67.9	3.0	−61.2	−9.8
High school	66.3	3.6	35.7	0.0
College	13.6	5.4	5.5	0.0
University	3.3	0.0	13.7	5.8
Married	−4.6	7.4	−8.1	−4.1
Have children	6.7	0.0	20.8	−4.1
Infectious and parasitic diseases	8.1	0.0	–	–
Cancer	8.2	0.0	−14.0	0.0
Mental disorders	−14.6	0.0	−17.8	0.0
Diseases of the nervous system	16.9	0.0	−20.8	8.4
Circulatory diseases	−18.8	0.0	−1.9	0.0
Respiratory diseases	0.0	0.0	–	–
Diseases of digestive system	−5.3	0.0	20.3	0.0
Genitourinary system diseases	−9.3	0.0	5.9	0.0
Musculoskeletal and soft-tissue disorders	−4.4	−4.9	14.2	−4.3
Congenital diseases	9.5	13.5	4.7	14.3
Zero earnings 1-year prior to application	9.5	−3.7	21.8	0.0
Zero earnings 2-years prior to application	11.6	−3.7	14.5	−13.1
Earnings 1-year prior to application	−18.7	−9.3	−9.5	−5.1
Earnings 2-years prior to application	−22.7	−5.8	−3.6	0.7
Unemployment rate	0.2	2.9	−19.0	−9.8
Duration on CPPD program	−9.7	−4.4	18.2	−9.6

Notes: Entries in the table are standardized differences between the treatment and comparison groups.

disorders, ill-defined causes and injuries and poisonings). Many of the other variables in the propensity score specification are not individually statistically significant, although they are jointly significant. The estimates for women in Table 6 resemble those for men. In particular, higher levels of educational attainment are associated with increases in the probability of participating in the CPPD-VR program and some health problems are associated with decreases in the probability of participating in the CPPD-VR, relative to the omitted category. Also like the estimates for males, most of the variables in the propensity score model are not individually statistically significant but the likelihood ratio test statistic for the model as a whole is quite large.

5.2 The average treatment effect on the treated for the CPPD-VR program

We present our estimates of the ATET for VR in Table 7 for men and Table 8 for women. We consider three outcomes (leaving the disability rolls, employment and substantial gainful employment) and present estimates of the ATET based on propensity score matching (kernel and local linear), inverse probability weighting and genetic matching. As we noted earlier, we present estimates for three ranges of the propensity score, with our preferred sample containing individuals with propensity scores that lie in [0.001, 0.40] for men and [0.001, 0.35] for women.

Table 6 Estimates of propensity score, logit model

	Men		Women	
	Coefficient estimate	Standard error	Coefficient estimate	Standard error
Age (at onset of disability)	0.519	0.761	−2.667**	1.067
Age^2	−0.009	0.021	0.076***	0.030
Age^3	0.000	0.000	−0.001***	0.000
Married	0.413	0.351	−0.427	0.331
Have children	0.334	0.321	0.188	0.339
[Less than high school]				
High school	2.673***	0.354	1.766***	0.428
Post-secondary	1.592***	0.492	1.311***	0.529
University degree	2.203***	0.544	2.169***	0.590
[Other]				
Infectious and parasitic diseases	−0.885	0.703	–	–
Cancer	−1.250*	0.729	−1.810**	0.877
Mental disorders	−1.580***	0.467	−1.613***	0.518
Diseases of the nervous system	0.804*	0.423	−1.806**	0.768
Circulatory diseases	0.021	0.548	0.175	0.702
Respiratory diseases	0.530	1.409	–	–
Diseases of digestive system	0.480	0.785	0.992	0.787
Genitourinary system diseases	−1.396	1.129	−0.071	1.046
Musculoskeletal and soft-tissue diseases	0.036	0.447	−0.398	0.496
Congenital diseases	−0.158	1.391	−0.509	1.217
Zero earnings 1-year prior to application	0.365	0.451	−0.097	0.524
Zero earnings 2-years prior to application	−0.273	0.505	−0.529	0.609
Earnings 1-year prior to application ('0,000 dollars)	0.000	0.000	0.000**	0.000
Earnings 1-year prior to pplication^2 ('000 dollars)	0.000	0.000	0.000**	0.000
Earnings 1-year prior to pplication^3 ('000 dollars)	0.000	0.000	0.000*	0.000
Earnings 2-years prior to application ('000 dollars)	0.000	0.000	0.000	0.000
Earnings 2-years prior to application^2 ('000 dollars)	0.000	0.000	0.000	0.000
Earnings 2-years prior to application^3 ('000 dollars)	0.000	0.000	0.000	0.000
Earnings 1-year prior to application × Earnings 2-years prior to application	0.000	0.000	0.000	0.000
Provincial unemployment rate	−0.231	3.602	−1.398	1.502
Provincial unemployment rate^2	0.095	0.418	0.119	0.145
Provincial unemployment rate^3	−0.006	0.016	−0.003	0.004
Duration on CPPD Program ('000 days)	−0.003	0.003	−0.003	0.003
Duration on CPPD Program^2 ('000 days)	0.000	0.000	0.000	0.000
Duration on CPPD Program^3 ('000 days)	0.000	0.000	0.000	0.000
Likelihood ratio test statistic	227.4 {<0.001}		163.9 {<0.001}	
Value of log likelihood	−198.15		−179.51	
Sample size	979		759	

Notes: ***denotes statistically significance at the 1 percent level. **denotes statistically significance at the 5 percent level. *denotes statistically significance at the 10 percent level. Square braces contain omitted reference category for dummy variables. Braces contain p-value for likelihood ratio test statistics.

Table 7 Estimates of the ATET of VR on individual outcomes, men

	Matching estimator		Inverse probability weighting	Genetic matching
	Kernel	Local linear		
	Sample restricted to propensity score values [0.001, 0.40]			
Leaving disability rolls	0.057	0.073	0.039	−0.036
	(0.082)	(0.113)	(0.076)	(0.096)
Gainful employment	0.062	0.070	0.030	−0.073
	(0.080)	(0.099)	(0.077)	(0.096)
Substantial gainful employment	0.025	0.035	0.001	−0.055
	(0.078)	(0.097)	(0.074)	(0.098)
Sample size	786	786	786	786
	Sample restricted to propensity score values [0.001, 0.35]			
Leaving disability rolls	0.014	0.015	0.003	−0.067
	(0.083)	(0.082)	(0.084)	(0.111)
Gainful employment	0.022	0.025	0.007	0.000
	(0.081)	(0.079)	(0.083)	(0.109)
Substantial gainful employment	0.009	0.026	0.013	−0.022
	(0.083)	(0.078)	(0.083)	(0.111)
Sample size	774	774	774	774
	Sample restricted to propensity score values [0.001, 0.50]			
Leaving disability rolls	0.033	0.044	0.018	0.000
	(0.082)	(0.085)	(0.072)	(0.101)
Gainful employment	0.024	0.027	−0.008	−0.017
	(0.081)	(0.088)	(0.073)	(0.097)
Substantial gainful employment	−0.008	−0.005	−0.033	−0.049
	(0.078)	(0.085)	(0.073)	(0.100)
Sample size	805	805	805	805

Notes: ATET denotes average treatment effect on the treated. Observations with propensity score values outside the range in square brackets are omitted from the sample used to estimate the ATET. Standard errors in parentheses. The kernel and local linear matching estimates were obtained with the Epanechnikov kernel, with bandwidths computed using cross-validation; see Table 1 for the bandwidths. Genetic matching estimates are based on single nearest neighbor matching with replacement. Standard errors for kernel and local linear matching and inverse probability weighting are obtained by bootstrapping with 1000 replications. Standard errors for genetic matching are based on Abadie and Imbens (2006).

The estimates for men in Table 7 indicate that VR has a relatively small effect on the outcomes we consider. Putting aside the genetic matching estimates for the moment, the other estimators suggest impacts on leaving the disability rolls and on gainful employment of about five or six percentage points. They indicate a smaller impact on substantial gainful employment of about two percentage points. None of these estimates differ statistically from zero, though of course the point estimates remain the preferred estimates. Mucking around with the common support does little to change the story.

The estimates for women in Table 8 differ from those for men: they are larger in magnitude and more often attain conventional levels of statistical significance. Looking across estimators, the estimates of the ATET for the leaving the disability rolls and gainful employment outcomes average a bit over ten percentage points, while the substantial gainful employment ATETs average around 16 percentage points. The former estimates do not attain statistical significance, while the latter do, sometimes at the five

Table 8 Estimates of the ATET of VR on individual outcomes, women

	Matching estimator		Inverse probability weighting	Genetic matching
	Kernel	Local linear		
Sample restricted to propensity score values [0.001, 0.35]				
Leaving disability rolls	0.104	0.104	0.095	0.122
	(0.086)	(0.084)	(0.080)	(0.111)
Gainful employment	0.110	0.109	0.108	0.143
	(0.082)	(0.081)	(0.078)	(0.104)
Substantial gainful employment	0.169**	0.158**	0.144*	0.184*
	(0.079)	(0.079)	(0.078)	(0.103)
Sample size	631	631	631	631
Sample restricted to propensity score values [0.001, 0.30]				
Leaving disability rolls	0.105	0.085	0.096	0.071
	(0.088)	(0.088)	(0.087)	(0.120)
Gainful employment	0.133	0.120	0.130	0.143
	(0.087)	(0.087)	(0.084)	(0.122)
Substantial gainful employment	0.180**	0.156*	0.159*	0.167
	(0.083)	(0.085)	(0.084)	(0.118)
Sample size	615	615	615	615
Sample restricted to propensity score values [0.001,0.40]				
Leaving disability rolls	0.109	0.093	0.093	0.077
	(0.086)	(0.081)	(0.078)	(0.105)
Gainful employment	0.147	0.137	0.138	0.154
	(0.085)	(0.076)	(0.074)	(0.099)
Substantial gainful employment	0.195**	0.175**	0.171**	0.192**
	(0.081)	(0.075)	(0.074)	(0.098)
Sample size	642	642	642	642

Notes: * denotes statistical significance at the 10 percent level; ** denotes statistical significance at the 5 percent level. See notes for Table 7.

percent level and sometimes at the 10 percent level. Changing the imposed common support region does not change the overall picture.

Looking across estimators in Tables 7 and 8, the genetic matching estimates end up outliers in both cases: on the high side for women and on the low side for men. The genetic matching estimates have the feature that they optimize balance and the feature that they build on a single nearest neighbor matching estimator that does not use as much of the information available in the comparison group as the other three estimators. For this reason, we tend to discount the estimates from the genetic matching estimator relative to the other three.

Substantively, the estimates reveal a large difference in impacts between men and women and, not unrelated, estimates for women large enough to cast doubt on the validity of our identification strategy. The estimates for women seem a bit large relative to a casual prior based on the nature of the treatment and of the participants' underlying conditions. To address our concerns about the magnitude of some of the estimates, and because we think it represents good empirical practice more generally, we turn now to an analysis of the sensitivity of our estimates to lingering selection on unobserved variables not accounted for by our choice of comparison group and our

conditioning variables. Later in the next section, we place our estimates in the broader context of the literature.

5.3 Sensitivity analysis and comparison with previous and related research

This section presents the results of a sensitivity analysis using the approach in Ichino et al. (2008). Their approach builds on earlier work by Rosenbaum and Rubin (1983) and Rosenbaum (1987), who consider the robustness of estimates of the ATET to assumptions with respect to a binary unobserved variable associated with both the treatment and outcome variables. Within the sensitivity analysis, the CIA holds when the variable is included and fails to hold when it is not included. Consequently, estimating the ATET including and excluding this variable from the conditioning set determines the sensitivity of the estimates of the ATET to the unobserved variable. Ichino et al. (2008) build on this framework and extend it so they can provide point estimates of the ATET under different assumptions about the distribution of the unobserved variable, rather than bounds as in the earlier literature.

We present the estimates from the sensitivity analysis in Tables 9 and 10. These tables present the estimates of the ATET by outcome measure. We present these estimates for various assumptions about the unobserved variable or confounder, denoted U. Each panel of estimates includes an estimate for the case with "no confounder", which means that the CIA holds and that no relevant variable has been excluded from the conditioning set and a "neutral confounder", which has a distribution that is calibrated so that it has no net effect on the untreated outcome and no effect on selection into treatment and so functions as a sort of placebo test. We also selected a few variables to simulate as confounders, which are listed in the rows of the tables. Unfortunately, the methodology we use restricts us to using discrete variables for the sensitivity analysis. We present estimates of the ATET as well as the quantities referred to as the "outcome" and "selection" effects in Ichino et al. (2008). The outcome and selection effects are odds ratios from logit models that estimate $P(Y = 1 \mid T = 0, X, U)$ and $P(T = 1 \mid X, U)$, where U is the confounder and the other variables are as previously defined. The outcome effect measures the effect of U on the untreated outcome controlling for observed covariates. The selection effect measures the effect of U on assignment to treatment controlling for observed covariates. The outcome and selection effects are used in the sensitivity analysis to benchmark how strong the confounder needs to be to change the substantive importance of the ATET or its statistical significance. The estimates we examine for the sensitivity analysis are the kernel matching estimates[3].

The sensitivity analysis for men in Table 9 indicates that the estimates of the ATET with the simulated confounders are the same as those with no unobserved confounding, but the precision of the estimates varies somewhat with the confounders. As with our estimates in Table 7, none of the estimates of the ATET are statistically significant. For women, we also see very robust findings, i.e., there is not a great deal of sensitivity of our estimates to the simulated confounder. For women, the ATET do not vary by the confounders we consider, but the precision of the estimates is affected somewhat for some confounders[4]. As we noted earlier, we are restricted to using binary variables in our sensitivity analysis. One explanation for our finding of no sensitivity is that the variables we selected are not strongly correlated with both treatment choice and outcomes. We feel that some of the continuous variables in our conditioning set, such as

Table 9 Sensitivity analysis, calibrated confounders for kernel matching estimate of ATET, men

	Outcome effect	Selection effect	ATET	SE
Leaving disability rolls				
No confounder	–	–	0.065	0.073
Neutral confounder	1.008	1.051	0.065	0.071
Confounders				
High school	1.703	3.508	0.065	0.081
College	3.194	2.213	0.065	0.090
University	1.566	1.325	0.065	0.083
Married	0.601	0.862	0.065	0.074
Mental disorders	0.750	0.812	0.065	0.084
Musculoskeletal disorders	0.660	0.972	0.065	0.081
Zero earnings 1-year prior to application	0.641	1.322	0.065	0.081
Zero earnings 2-years prior to application	0.713	1.515	0.065	0.074
Gainful employment				
No confounder	–	–	0.074	0.081
Neutral confounder	1.011	1.028	0.074	0.084
Confounders				
High school	1.981	3.640	0.074	0.071
College	2.990	2.175	0.074	0.083
University	1.247	1.278	0.074	0.072
Married	0.560	0.868	0.074	0.078
Mental disorders	0.862	0.805	0.074	0.082
Musculoskeletal disorders	0.599	0.994	0.074	0.083
Zero earnings 1-year prior to application	0.554	1.304	0.074	0.071
Zero earnings 2-years prior to application	0.647	1.509	0.074	0.077
Substantial gainful employment				
No confounder	–	–	0.051	0.075
Neutral confounder	1.017		0.051	0.069
Confounders				
High school	1.852	3.634	0.051	0.069
College	2.946	2.196	0.051	0.065
University	1.352	1.298	0.051	0.087
Married	0.569	0.872	0.051	0.062
Mental disorders	0.085	0.797	0.051	0.073
Musculoskeletal disorders	0.639	0.974	0.051	0.069
Zero earnings 1-year prior to application	0.510	1.300	0.051	0.087
Zero earnings 2-years prior to application	0.593	1.532	0.051	0.070

Notes: ATET denotes average treatment effect on the treated; SE denotes standard error. The outcome and selection effects are odds ratios. The rows in the tables denote the confounding variable, the first two rows in each panel of estimates show the ATET with no cofounding factor and with a neutral confounder. The kernel and local linear matching estimates were obtained with the Epanechnikov kernel, with bandwidths computed using cross-validation. All standard errors are obtained by bootstrapping with 1000 replications.

pre-program earnings and spell duration on the CPPD program, would be better candidates for confounders since they are proxies for otherwise unobserved variables. Still, we find a remarkable lack of sensitivity.

Table 10 Sensitivity analysis, calibrated confounders for kernel matching estimate of ATET, women

	Outcome effect	Selection effect	ATET	SE
	Leaving disability rolls			
No confounder	–	–	0.103	0.068
Neutral confounder	1.018	1.034	0.103	0.078
Confounders				
High school	1.295	2.235	0.103	0.067
College	1.464	1.219	0.103	0.078
University	1.358	1.703	0.103	0.077
Married	0.854	0.885	0.103	0.074
Mental disorders	1.045	0.727	0.103	0.069
Musculoskeletal disorders	0.652	1.433	0.103	0.088
Zero earnings 1-year prior to application	0.468	1.712	0.103	0.090
Zero earnings 2-years prior to application	1.064	1.496	0.103	0.082
	Gainful employment			
No confounder	–	–	0.141**	0.069
Neutral confounder	1.016	1.043	0.141	0.089
Confounders				
High school	1.166	2.232	0.141**	0.069
College	1.624	1.198	0.141**	0.070
University	1.434	1.689	0.141**	0.068
Married	1.031	0.889	0.141**	0.063
Mental disorders	0.904	0.707	0.141**	0.07
Musculoskeletal disorders	0.569	1.456	0.141	0.088
Zero earnings 1-year prior to application	0.288	1.747	0.141*	0.080
Zero earnings 2-years prior to application	0.704	1.466	0.141**	0.070
	Substantial gainful employment			
No confounder	–	–	0.200***	0.066
Neutral confounder	1.019	1.031	0.200**	0.079
Confounders				
High school	1.176	2.202	0.200**	0.082
College	1.551	1.214	0.200**	0.077
University	1.592	1.678	0.200**	0.081
Married	0.961	0.899	0.200**	0.077
Mental disorders	0.998	0.716	0.200**	0.075
Musculoskeletal disorders	0.529	1.431	0.200***	0.069
Zero earnings 1-year prior to application	0.241	1.703	0.200***	0.066
Zero earnings 2-years prior to application	0.643	1.455	0.200**	0.079

Notes: See the notes to Table 9.

Our estimates indicate that VR has a larger effect on the likelihood of leaving the disability rolls or finding employment for women. Our findings are not unlike the findings from the literature examining active labor market programs that has generally found that such programs have a larger effect on the labor market outcomes of women than those of men (e.g., Heckman et al. 1999). However, the large difference in the size of the estimates for men and women, as well as the absolute size of the estimated ATET for women, make

us somewhat cautious about drawing strong conclusions about the improvement in employment outcomes of women after receiving VR. While we use a large and very relevant set of conditioning variables motivated by theory, institutions and the prior literature, in combination with a flexible functional form, some selection bias may remain in our estimates. We would have liked to condition on a measure of the severity of each individual's disability, on some measure of ability (i.e. a test score), on measures of non-cognitive skills, and so on. We think that using pre-disability labor market outcomes captures much of what we would want from these variables, but it may not go all the way. We also worry that duration on CPPD may not fully capture the variation in human capital depreciation across individuals. Alternatively, as we noted earlier, drawing our comparison group from the reassessment file may not accomplish what we intend it to accomplish if individuals in the reassessment file differ in, say, unobserved motivation to return to work.

Most of the previous literature looking at VR has examined programs with different rehabilitation strategies as well as different client bases and many of these studies may also be sensitive to specification errors or contaminated by other problems, as they sometimes do not use very transparent identification strategies. However, our results do have some parallels in some more recent empirical work studying VR programs with more clearly defined identification strategies. For example, Aakvik et al. (2005) find an (imprecisely estimated) increase in the employment of women in their study of the effects of the Norwegian VR program, though it is not as large as ours. Aakvik (2001) also finds positive (but imprecisely estimated) effects of the Norwegian VR program on employment for a pooled sample of men and women, with a point estimate of 6.3 percentage points, but interprets his estimates with caution due to lingering worries about selection bias resulting from his bounding analysis. We also estimated the ATET using an IPW estimator with a pooled sample of men and women and obtained estimates similar to those in Aakvik (2001). More specifically, we obtained an ATET of 0.061 for the gainful employment outcome and 0.052 for the substantial gainful employment outcome, but these estimates did not statistically differ from zero.

Our estimates of the effect of VR are also quite interesting in comparison to estimates from studies looking at financial and non-financial incentives to increase the attachment of disability beneficiaries to the labor market. Campolieti and Riddell (2012) found that the introduction of the CPPD earnings disregard (similar to the benefit off-set in the SSDI program in the U.S.) increased the employment of disability beneficiaries. Campolieti and Riddell's preferred difference-in-difference estimates show that the introduction of the earnings disregard and automatic reinstatement option increase the employment of men by 5.1 percentage points and women by 9.5 percentage points, relative to the Quebec Pension Plan Disability program. One concern about the introduction of such incentives is the increased uptake of benefits, which is also referred to as the induced entry effect (e.g. Hoynes and Moffitt 1999). However, Campolieti and Riddell (2012) did not find any evidence of the increased receipt of disability benefits in their analysis of flows onto the disability rolls. Kostøl and Mogstad (2013) looked at the introduction of a benefit offset in Norway, which is very similar to the benefit offset used in the SSDI program, exploiting a discontinuity created by the eligibility rules for these new incentives, and found that there was a five to six percentage point increase in employment after the introduction of the benefit offset. They also found that the

benefit offset could generate program savings of about 3.5 to 5.0 percent of costs of SSDI benefits.

6. Cost savings to the CPPD program from the VR program

This section presents the results from a crude, back-of-the-envelope cost-benefit analysis from the narrow perspective of the CPPD program. Alternatively, a cost-benefit analysis could be undertaken from the perspective of persons with disabilities, taxpayers and society as a whole. Our analysis from the perspective of the CPPD program reflects the net (of VR programs costs) savings in transfers from the program to persons with disabilities associated with providing VR and not a welfare gain or loss to society.

We measure the benefits of the VR program as the CPPD payments that the program saves over the remainder of a claimants working life when they leave the disability rolls combined with the additional (involuntary) contributions they make as a result of working. The costs consist of the expenditures on VR for the treatment group. We used a real discount rate of 3.9 percent in our base-case computations, which is the mean of the interest rate on real return government of Canada bonds from 1997 to 2001. We conducted a sensitivity analysis with a lower bound discount rate of 1.9 percent and an upper bound of 5.9 percent.

To keep things very simple, we choose a discounting horizon equal to the difference between 65, the CPP regular retirement age when all disability benefits roll over into retirement benefits, and 47, the average age at which individuals leave the disability rolls in the comparison group; this difference equals 18 years. Our calculations implicitly assume that individuals who leave the disability rolls do not return at some later point. This assumption likely makes our estimates an upper bound on the effects of the VR program on CPPD transfers to individuals.

In order to incorporate the higher probability that individuals in the treatment group leave the disability rolls, we compute

$$ ENB = q_{leave} \times \left[\sum_{t=0}^{T} \frac{B_t}{(1+r)^t} \right] - C \tag{4} $$

where ENB denotes expected net benefits, q_{leave} is the probability of leaving the disability rolls, B is benefit payments saved, C is the per capita cost of VR and r is the discount rate. The corresponding formula for the comparison group simply omits the C. We computed the q_{leave} probabilities for the treatment group based on the IPW estimates of the ATET (i.e., the proportion in the comparison group plus the ATET), while the proportion leaving the CPPD program is used as the q_{leave} probability for the comparison group. The specific values that we plug into the formula along with additional details about the calculations appear in Table 11 and its associated table notes.

Table 12 presents the net expected benefits (per person) for the treatment and comparison groups. For the base case (a discount rate of 3.9 percent), our estimates indicate that the net impact of the CPPD-VR program (the ENB for the treatment group less the ENB for the comparison group) for women equals $4,538 per client. During the 1998–1999 fiscal year the mean annual CPPD benefit payment to an individual was about $8,968, so our preferred estimate of the cost savings is about 0.5 years of the

Table 11 Information for cost-benefit analysis

	Probability of benefits ceasing	Discounting horizon (years)	Mean benefits (dollars)	Mean VR cost (dollars)
		Women		
Treatment Group	0.373	18	7,908	5,783
Comparison Group	0.278	18	8,994	0
		Men		
Treatment Group	0.412	18	7,567	6,376
Comparison Group	0.373	18	8,110	0

Notes: The probability of ceasing benefits is the mean for the comparison group; for the treatment group the probability of ceasing benefits is computed as the mean for the comparison group plus the IPW estimate of the ATET in Tables 7 and 8 based on the sample that restricts the propensity score to [0.001, 0.40] for men and [0.001, 0.35] for women. The discounting horizon is computed as the retirement age (65) minus the average age that benefits cease for the comparison group (47). Payroll tax contributions to the CPP are based on the average pre-CPPD earnings multiplied by 0.099, which is the contribution rate since 2003.

average benefit payment per VR client. The estimates are not overly sensitive to the discount rate assumption, with the estimate increasing to $5,687 (0.63 years of mean benefits) with a discount rate of 1.9 percent and falling to $3,637 (0.41 years of mean benefits) with a discount rate of 5.9 percent.

In sharp contrast to the value obtained for women, the present value of the net benefits of the CPPD-VR program for men turns out negative. For the base case discount rate it equals about -$2,728, which indicates that the CPPD-VR program does not generate any cost savings for men. The estimated savings (or lack thereof) associated with the CPPD-VR program for men also do not vary very much with the assumed discount rate.

Our estimates indicate cost savings for the CPPD-VR program from serving women but not from serving men. This suggests the potential value of targeting VR services differentially toward female claimants, as well as the value of additional research designed to account for the difference in impacts between men and women and to identify other subgroups defined by observed characteristics that experience differentially larger impacts from the VR program.

7. Conclusion

This paper estimates the effect of the CPPD-VR program on the labor market outcomes of disability beneficiaries. Our identification strategy rests on the selection of a comparison group of reassessed claimants that we think reduces the amount of selection we face relative to a broader comparison group, combined with a "selection on

Table 12 Expected net benefits to the CPPD program for individuals who leave the disability rolls

	Men	Women
	3.9% discount rate	
Net Impact of CPPD-VR	−1296	4032
	1.9% discount rate	
Net Impact of CPPD-VR	−1128	5094
	5.9% discount rate	
Net Impact of CPPD-VR	−1519	3378

Notes: All entries are in dollars. The net impact is computed as the difference in the present value of the net benefits in the treatment group minus the present value of the net benefits in the comparison group. The base discount rate of 3.9 percent is the average return on real return Canadian bonds between Jan. 1, 1997 and Dec. 31, 2001. Table 11 and the notes to that table provide further details on inputs into the calculation.

observed variables" assumption rendered (somewhat) plausible by our data on CPPD spell duration, pre-disability labor market outcomes and demographics. That variation in the number of CPPD-VR slots does not depend on local economic conditions also adds to the credibility of our analysis, as do the flexible specification we employ in our propensity score model, our comparison of multiple matching and weighting estimators and our additional analysis of the sensitivity of our estimates to the presence of unobserved confounders. We argue that we have done about the best we can with the data available to us and the institutions generating variation in treatment receipt.

Our estimates suggest that CPPD-VR improved the labor market outcomes of women enough to pass a cost-benefit test from the perspective of the program. We cannot say the same for the men. Though some of the estimates for women attain conventional levels of statistical significance, overall our estimates lack the precision required to detect modest but not trivial effects. These issues of precision follow directly from the sample sizes available to us combined with our reliance on semi-parametric and non-parametric estimation strategies. Our estimates comport with others in the literature in a broad sense, but remain a bit too high for the women relative to our prior. Despite this, the estimates for women (and for men) demonstrate very little sensitivity to unobserved confounders.

Finally, perhaps the most important lesson from our paper concerns the evidence it provides regarding the ability of researchers to perform really credible analyses of VR program impacts. In our context, as in many others involving programs for persons with disabilities, we lack the data and/or controlled variation in treatment status required to do a very compelling analysis. We provided a list of variables we would have liked to condition on above. Adding those variables to the administrative data would cost relatively little and would allow many other interesting analyses as well. Alternatively, those who run the CPPD program, and their peers who run similar programs elsewhere, could do more to introduce well-defined mechanisms that vary treatment status, such as random assignment, or the use of numerical readiness scores to assign treatment (thus allowing a discontinuity design) or randomizing the roll-out of new programs across program offices and so on. In the absence of additional data and/or better variation, policymakers and researches must settle for making an evidentiary meal with whatever stray morsels the data and variation cupboard contains, as we ourselves have done in this paper. Perhaps persons with disabilities deserve better?

Endnotes

[1]These income thresholds are determined by the CPP disability program and were $8,937 in 1998, $9,020 in 1999, $9,155 in 2000 and $9,300 in 2001.

[2]In Canada, labor market earnings are reported on the T4 tax form, which we have access to in the ROEMF. The T4 form in Canada is like the W2 form in the U.S.

[3]The sensitivity analysis uses the ado file created by Nannicini (2007) and thus the kernel matching estimator in the propensity score matching software developed by Becker and Ichino (2002). We modified their ado file so that it would use the bandwidth we selected via cross-validation and the Epanechnikov kernel rather than the default choices the program otherwise imposes. The Becker and Ichino (2002) kernel matching estimator is not coded in the same manner as its counterpart in the psmatch2 ado file we use to obtain the results presented in Tables 7 and 8. As a result, there are some minor differences between the ATET estimates from the main analysis and the baseline estimates in the sensitivity analysis.

[4]We also tried some other controls for health problems and household characteristics and found similar findings to those we report in the tables.

Author details
[1]Department of Management, University of Toronto (Scarborough Campus), Toronto, Canada. [2]Centre for Industrial Relations, University of Toronto, Toronto, Canada. [3]Department of Economics and Centre for Industrial Relations, University of Toronto, Toronto, Canada. [4]Department of Economics, University of Michigan, Ann Arbor, USA. [5]National Bureau of Economic Research and Institute for the Study of Labor (IZA), 238 Lorch Hall, 611 Tappan Street, Ann Arbor, MI 48109-1220, USA.

References
Aakvik A (2001) Bounding a matching estimator: the case of a Norwegian training program. Oxf Bull Econ Stat 63:115–143
Aakvik A, Heckman J, Vytlacil E (2005) Estimating treatment effects for discrete outcomes when responses to treatment vary: an application to Norwegian vocational rehabilitation programs. J Econ 125:15–51
Abadie A, Imbens G (2006) Large sample properties of matching estimators for average treatment effects. Econometrica 74:235–267
Allingham R, Hyatt D (1995) Measuring the impact of vocational rehabilitation on the probability of post-injury return to work. In: Thomason T, Chaykowski R (ed) Research in workers' Compensation. Queen's University IRC Press, Kingston, Ontario
Anderson R, James MK, Miller ME, Worley AS, Longino CF (1998) The timing of change: patterns in transitions in functional status among elderly persons. J Gerontol 53B:S17–S27
Andersson F, Holzer H, Lane J, Rosenblum D, Smith JA (2013) Does federally-funded job training work? Nonexperimental estimates of WIA training impacts using longitudinal data on workers and firms.. NBER working paper #19446
Becker SO, Ichino A (2002) Estimation of average treatment effects based on propensity scores. Stata J 2:358–377
Beckett LA, Brock DB, Lemke JH, Mendes de Leon CF, Guralnik JM, Fillenbaum GF, Branch LG, Wetle TL, Evans DA (1996) Analysis of change in self-reported physical function among older persons in four population studies. Am J Epidemiol 143:766–778
Black DA, Smith JA (2004) How robust is the evidence on the effects of college quality? Evidence from matching. J Econ 121:99–124
Bound J, Burkhauser RV (1999) Economic analysis of programs targeted on people with disabilities. In: Ashenfelter O, Card D (ed) Handbook of Labor Economics, Volume 3. North-Holland, Amsterdam
Busso M, DiNardo J, McCrary J (2013) New evidence on the finite sample properties of propensity score matching and reweighting estimators. Rev Econ Stat. in press
Campolieti M (2002) Moral hazard and disability insurance: on the incidence of hard-to-diagnose medical conditions in the Canada/Quebec Pension Plan disability program. Can Public Pol 28:419–421
Campolieti M (2009) Worker adaptation and the desire for accommodations after the onset of a disability. Ind Relat 48:329–349
Campolieti M, Hyatt D (2011) Does vocational rehabilitation have much of an impact on helping injured workers return to work? In: Taras D, Williams K (ed) Disability and the Workplace. National Institute for Disability Management and Research, Port Alberni, BC
Campolieti M, Riddell C (2012) Disability policy and the labor market: evidence from a natural experiment in Canada, 1998–2006. J Public Econ 96:306–316
Crump RK, Hotz VJ, Imbens GW, Miknik OA (2009) Dealing with limited overlap in the estimation of average treatment effects. Biometrika 96:187–199
Currie J, Madrian B (1999) Health, Health insurance and the labor market. In: Ashenfelter O, Card D (ed) Handbook of Labor Economics, Volume 3. North-Holland, Amsterdam
Daly MC, Bound J (1996) Worker adaptation and employer accommodation following the onset of a disability. J Gerontol 51B:S53–S60
Dean D, Dolan R (1991a) Fixed effects estimates of earnings impacts for the vocational rehabilitation program. J Hum Resour 26:381–391
Dean D, Dolan R (1991b) Assessing the role of vocational rehabilitation in disability policy. J Policy Anal Manage 10:568–587
Dean D, Pepper J, Schimdt R, Stern S (2013a) The effects of vocational rehabilitation for people with cognitive impairments. http://people.virginia.edu/~sns5r/resint/vocrehstf/vocrehci.pdf. Accessed 3 March 2014

Dean D, Pepper J, Schimdt R, Stern S (2013b) The effects of vocational rehabilitation for people with mental illness. http://people.virginia.edu/~sns5r/resint/vocrehstf/vocrehmi.pdf. Accessed 3 March 2014

Dehejia RH, Wahba S (1999) Causal effects in non-experimental studies: reevaluating the evaluation of training programs. J Am Stat Assoc 94:1053–1062

Diamond A, Sekhon J (2013) Genetic matching for estimating causal effects: a general multivariate method for achieving balance in observational studies. Rev Econ Stat 95:932–945

Dolton P, Smith JA (2011) The impact of the UK new deal for lone parents on benefit receipt.. IZA Working Paper #5491

Gardner J (1988) Improving vocational rehabilitation outcomes for earlier intervention. Workers' Compensation Research Institute, Cambridge, MA

Grossman M (1972) On the concept of health capital and the demand for health. J Polit Econ 80:223–255

Hahn J (1998) On the role of the propensity score in efficient semiparametric estimation of average treatment effects. Econometrica 66:315–331

Heckman J, Robb R (1985) Alternative methods for evaluating the impact of interventions. In: Heckman J, Singer B (ed) Longitudinal Analysis of Labor Market Data. Cambridge University Press, Cambridge

Heckman J, Ichimura H, Todd P (1997) Matching as an econometric evaluation estimator: evidence from a job training programme. Rev Econ Stud 64:605–654

Heckman J, LaLonde R, Smith JA (1999) The economics and econometrics of active labor market programs. In: Ashenfelter O, Card D (ed) Handbook of Labor Economics, Volume 3A. Elsevier, Amsterdam

Heinrich C, Mueser P, Troske K, Jeon K, Kahvecioglu D (2013) Do public employment and training programs work? IZA J Labor Econ 2:6

Hoynes H, Moffitt R (1999) Tax rates and work incentives in the social security disability insurance program. Natl Tax J 52:623–654

Ichino A, Mealli F, Nannicini T (2008) From temporary help jobs to permanent employment: what can we learn from matching estimators and their sensitivity? J Appl Econ 23:305–327

Kostøl AR, Mogstad M (2013) How financial incentives induce disability beneficiaries to return to work. Am Econ Rev. in press

Lee W (2013) Propensity score matching and variations on the balancing test. Empir Econ 44:47–80

Leuven E, Sianesi B (2003) PSMATCH2: Stata module to perform full Mahalanobis and propensity score matching, common support graphing, and covariate imbalance testing. http://ideas.repec.org/c/boc/bocode/s432001.html. Version 3.0.0

Meyer BD, Mok W (2013) Disability, earnings, income and consumption. NBER Working Paper #18869

Nannicini T (2007) Simulation-based sensitivity analysis for matching estimators. Stata J 7:334–350

Rosenbaum P (1987) Sensitivity analysis to certain permutation inferences in matched observational studies. Biometrika 74:13–26

Rosenbaum P, Rubin DB (1983) The central role of the propensity score in observational studies for causal effects. Biometrika 70:41–55

Rupp K, Scott CG (1998) Determinants of duration on the disability rolls and program trends. In: Rupp K, Stapleton DC (ed) Growth in Disability Benefits: Explanations and Policy Implications. Upjohn Institute for Employment Research, Kalamazoo, MI

Sekhon J (2011) Matching: multivariate and propensity score matching with automated balance search. J Stat Softw 42:1–52

Skaburskis A, Collignon F (1991) Cost-effectiveness analysis of vocational rehabilitation services. Can J Program Eval 6:1–23

Smith JA, Todd P (2005) Does matching overcome LaLonde's critique of nonexperimental estimators? J Econ 125:305–353

Social Development Canada (2004) Evaluation of the Canada Pension Plan Disability Vocational Rehabilitation Program. Final Report. Audit and Evaluation Directorate, Strategic Direction, Social Development Canada, Ottawa

Subcommittee on the Status of Persons with Disabilities (2004) Final Report. Queen's Printer, Ottawa. Final Report. Queen's Printer, Ottawa

Todd P (2008) Matching estimators. In: Durlauf SN, Blume LE (ed) The New Palgrave Dictionary of Economics, 2nd edition. Palgrave Macmillan, New York

U.S. General Accounting Office (1987) Social Security: little success in rehabilitating disabled beneficiaries. U.S. Government Printing Office, Washington, DC

U.S. General Accounting Office (1994) Social Security disability rolls keep growing, while explanations remain elusive. U.S. Government Printing Office, Washington, DC

U.S. Social Security Administration (2004) Ticket to Work evaluation report. http://www.ssa.gov/disabilityresearch/ttw/ttw_report.htm. Accessed 3 March 2014

Wittenburg D, Mann DR, Thompkins A (2013) The disability system and programs to promote employment for people with disabilities. IZA J Labor Pol 2:4