

**The Performance of Performance-based Contracting in Human Services:  
A Quasi-experiment**

Jiahuan Lu  
Ph.D. Candidate  
Maryland School of Public Policy  
University of Maryland  
College Park, MD 20742  
Tel: 240-898-8132  
Email: [jhlu@umd.edu](mailto:jhlu@umd.edu)

\*\*\* Draft \*\*\*

**Paper Presented at the 2013 APPAM Fall Research Conference, Washington, DC,  
November 7-9, 2013**

# **The Performance of Performance-based Contracting in Human Services: A Quasi-experiment**

## **Abstract**

Performance-based contracting (PBC) is becoming increasingly attractive to public human service agencies. By attaching contract compensation to contractors' performance achievement, PBC is expected to encourage quality services, better outcomes, and less monitoring. However, current empirical evidence on the effectiveness of PBC is still limited and mixed. This paper, based on a case study of the Indiana vocational rehabilitation program, employs a quasi-experimental design to evaluate the effectiveness of PBC on individual employment outcomes over a period 2004 – 2009, with Michigan program as a control. After using propensity score matching and difference-in-differences regressions to control for the imbalances between the two states, this paper finds PBC is significantly effective in promoting better employment results and shorter time-to-employment, two measured performance incentives. The PBC impact on unmeasured employment quality, demonstrated by working hours and wages, is trivial. This paper further suggests introducing relational contracting as a supplement when using PBC in the purchase of human services.

## **Introduction**

Within the “contracting regime” (Smith and Lipsky 1993) and the “hollow state” (Milward and Provan 2000), there is an increased reliance on contracting in human service delivery. Third-party actors have frequently been involved in the production of a wide range of services on behalf of different levels of governments. This significant explosion of contracting has fundamentally redefined the features of the American governance system. In a political sense, contractors become a buffer between the state and citizens, representing the missions and goals of grand social programs (Smith and Lipsky 1993). In a managerial sense, since government programs are dependent on contract operation, the performance of government has become largely contingent on contractors (Frederickson and Frederickson 2006; Kettl 2002). Put together, both raise the critical issue of contracting management.

However, managing contracting is very different from managing service production within traditional government bureaucracies. The reliance on contracting in public management discourse represents a significant shift away from a vertical, authority-based model to a horizontal, negotiation-driven model (Cooper 2003). The basic administrative responsibility thus becomes arranging and overseeing networks rather than managing hierarchies (Milward and Provan 2000). As a result, a central puzzle for public managers, as Kettl (2002, 493) summarizes, is that “[t]hey are responsible for ensuring high-quality results in programs that they do not directly control.” Unfortunately, public managers are not well prepared to confront this challenge (Johnston and Romzek 1999; Kettl 1993; Van Slyke 2003). To address this gap, public management scholarship in the last two decades has been marked by a surge of experimentation with various capacity-building activities (Amirkhanyan 2010; Brown and Potoski 2004; Hefetz

and Warner 2004; Lambright 2009; Romzek and Johnston 2005). The rise of performance-based contracting (PBC) represents one of the recent efforts.

PBC, by including performance measures into contract specifications and attaching contract compensation to these measures, is believed to enable quality services, better outcomes, and less monitoring. Currently, PBC is being increasingly used as a preferred contracting approach over the traditional fee-for-service (FFS) method in a variety of human service areas (Brucker and Stewart 2011; Heinrich and Choi 2007; McBeath and Meezan 2010; McLellan, Kemp, Brooks and Carise 2008; O'Brien and Revell 2005). However, despite the burgeoning popularity, the effectiveness of PBC compared with FFS in the purchase of human services is still unclear. Actually, human services are not always considered compatible with performance measurement, largely due to their ambiguous performance and high provider discretion (Hasenfeld 1983; Lipsky 1980). Relying on imperfect surrogate measures leaves service contractors room to gaming, while high provider discretion helps contractors gain these potential benefits (Bevan and Hood 2006; Courty and Marschke 2004; Heckman, Heinrich and Smith 2002). Therefore, human services bring challenges to PBC and make it at the risk of “rewarding A, while hoping for B” (Kerr 1975).

Indeed, current documented evidence on the effectiveness of PBC in human services is limited and mixed. In substance abuse treatment programs, Commons, McGuire, and Riordan (1997) observed significant reductions in clients' drug use and service costs after contractors were evaluated on service effectiveness, efficiency, and special populations served in Maine. This finding was largely in doubt later because it failed to discuss the unintended effects triggered by PBC such as client selection (Shen 2003) and misreporting and cheating (Lu 1999). Brucker and Stewart (2011) reexamined Maine's experience and concluded that PBC had no

positive effect on program performance such as time to treatment, level of client participation, length of stay, and completion of treatment. In employment services, the use of PBC in the programs funded by the Job Training Partnership Act has been found very controversial (Barnow 2000; Heckman, Heinrich and Smith 2002). Heinrich and Choi (2007) reported contractors under PBC in the Wisconsin Works program did respond to performance incentives, but insufficient administrative capacities undermined the effectiveness of PBC. In a local welfare-to-work training program, Dias and Maynard-Moody (2006) noticed when contract payment was linked to job placement and profit quotas, PBC created considerable tensions between managers and front-line workers on the importance of meeting performance goals versus meeting client needs, which further led to negative program operation and service outcomes.

This study, relying on a quasi-experiment, seeks to add to this line of research on the effectiveness of PBC in human services. Based on a case study of the Indiana vocational rehabilitation (VR) program, this paper employs administrative data to examine the impact of PBC on individual employment outcomes over a period 2004 - 2009<sup>1</sup>, with Michigan VR program as a control. During this period, Indiana, as the treatment group, converted from FFS to PBC in its purchase of VR placement services in the end of 2006, while Michigan, the neighbor north of Indiana, continued using FFS approach. After using propensity score matching and difference-in-differences regressions to control for both the observed and unobserved imbalances between the two states, this paper finds PBC is significantly effective in promoting better employment results and shorter time-to-employment, two measured performance incentives. The impact on unmeasured employment areas, demonstrated by working hours and wages, is trivial. This finding warns the dynamics of introducing performance management to human service

---

<sup>1</sup> All the years in this research refer to fiscal years.

contracting and further raises the policy concern of how to take full advantage of PBC. In addition to technically restructure PBC systems, the paper emphasizes the supplementary role of relational contracting. Generally, the more provider discretion and ambiguous service performance involved in human service delivery, the less effectiveness PBC as a formal control mechanism could produce, which leaves room for relational contracting to fit in..

The paper proceeds as follows. It begins with a brief discussion of the background of the rise of PBC in human services. This is followed by a theoretical discussion of two contract arrangements (behavior-oriented and outcome-oriented contracts) and their applicability in human service provision. After that, vocational rehabilitation programs are introduced as a policy field, with details of the PBC model in the purchase of placement services in Indiana. The detailed design of the quasi-experiment is discussed in section four. Section five presents the results of the quantitative analysis. Finally, I elaborate on the policy implications and limitations of the findings.

## **I. The Rise of PBC in Human Services**

This paper defines PBC in a loose way as an “umbrella” term: PBC incorporates performance measures into contract specification and makes contract compensations (such as payment, extension, and renewal) fully or partially contingent on performance achievements<sup>2</sup>. When using PBC, public managers only specify the desired end results of contracted services, leaving contractors substantial freedom to prescribe service delivery methods and use of funds. Theoretically, PBC is expected to promote better outcomes, quality services, and acquisition efficiency. First, by making contract compensation attached to performance achievement, PBC

---

<sup>2</sup> PBC may also be referred to as result-based contracting, performance-based acquisition, and result-based funding in different contexts.

draws contractors' attention toward the results of service delivery, rather than the delivery of services per se. In this way, PBC encourages outcome improvement. Second, by leaving contractors much freedom in the service process, PBC expect contractors to use such discretion to design best practice and use quality and innovative services to match client needs, rather than dealing with extensive administrative reporting and paperwork, which would again enhance service outcomes. Combining these, PBC promises greater government acquisition efficiency. Under PBC, only desired outcomes would be reimbursed, which maximizes the productivity of administrative resources.

In essence, PBC in human service provision stands for a marriage of service contracting with performance management, two prevalent managerial tools in contemporary public administrative narrative. On one hand, human service contracting has been a common and desired practice for decades, with its historical roots back to the colonial period (Smith and Lipsky 1993). Along with the widespread use of contracting, public managers often lack sufficient capacity to use contracting effectively (Johnston and Romzek 1999; Kettl 1993; Van Slyke 2003). In this vein, PBC, by introducing performance measures into contracting management, can be seen as an endeavor in helping address this capacity deficit. On the other hand, largely prompted by the government reinventing movement in the early 1990s, governments at all levels started to link resource allocation with performance measures and establish a variety of pay-for-performance systems (Behn 2003; Hatry and Wholey 1999; Kravchuk and Schack 1996). At the outset, these efforts were mostly run within traditional government domain. However, as more and more indirect government tools are introduced into the governance system (Salamon, 2002), it seems inevitable to witness the expansion of

performance elements to the management of indirect tools, forming a relatively comprehensive performance management system. PBC thus becomes an indispensable part therein.

Traditionally, human service contracting is run on a fee-for-service (FFS) basis, with an orientation on service input and process. Therefore, in FFS contracts, there are often clear specifications of input standards and service delivery procedures, such as detailed equipment and technologies to be used and a defined amount of time and labor required. After services are delivered, contractors are reimbursed based on unit of service delivered. Compared with FFS, conceptually, PBC represents several changes in the landscape of human service contracting. First, PBC changes the contract specification method, from a design specification of input and process to a performance specification of output, quality, and outcome (Martin, 2005). In this way, PBC presents new managerial responsibilities in specifying outcomes, designing incentives, and evaluating results. Third, the same as other performance management strategies, PBC implies a change in accountability relationship, with an increasing attention to accountability for results. It represents a switch away from hierarchical accountability with input and process orientations toward professional accountability which allows for the exercise of professional discretion and expertise (Romzek and Johnston 2005).

## **II. Theoretical Framework**

The same as previous literature on government contracting (Kettl 1993; Milward and Provan 2000; Romzek and Johnston 2005), this research puts the discussion of contract design in a principal-agent framework, where government (the principal) relies on contractors (the agent) to deliver services and achieve policy goals. Based on the assumptions of goal conflicts and information asymmetry between the principal and the agent, the agency theory highlights the agency problems, i.e., the principal is subject to the agent's self-serving opportunistic behaviors,

including adverse selection and moral hazard. Given this, the managerial implication of the agency theory focuses on the design of appropriate control mechanisms to guide the distribution of risk and uncertainty between the principal and the agent.

When organizational control is seen as a problem of information flow (Ouchi & Maguire, 1975), the design of control mechanisms primarily involves two dimensions: (1) task programmability – the principal’s capability to precisely define the means-ends relationships for certain tasks, and (2) outcome measurability – the principal’s capacity to specify the various aspects of task outcomes in a comprehensive and quantifiable manner. The focus of control, therefore, can be on either the behaviors of the agent or the outcomes of those behaviors (Ouchi, 1979; Thompson, 1967). Generally, when certainty regarding causation between means and ends is high, control strategies can be reflected in a high level of monitoring and direction in agent activities, with performance evaluation often on job inputs and process. If outcome measurability is high, the principal would prefer outcome-based control strategies, under which compensation schemes are attached to outcome measures and monitoring of employee activities becomes relatively less necessary.

Arrow (1964) defines the design of control strategies as the choice of operating rules and the choice of enforcement rules to support the operating rules. If an organization operates “as a nexus for a set of contracting relationships among individuals” (Jensen and Meckling, 1976, 310), then the design of the optimal contract arrangement governing the principal-agent relationship constitutes the enforcement rule to facilitate contract implementation. In accordance with two types of organizational control, there are two major contract alternatives: behavior-based and outcome-based contracts. The choice of a contract type is thus a function of task programmability and outcome measurability. The key in structuring the contractual relationship,

writes Eisenhardt (1989, 61), is “the trade-off between (a) the cost of measuring behavior and (b) the cost of measuring outcomes and transferring risk to the agent.”

**Figure 1. The Determinants of Contract Arrangements**

		<b>Task Programmability</b>	
		Low	High
<b>Outcome Measurability</b>	Low	Cell 1	Cell 2 <b>Behavior-based Contract</b>
	High	Cell 3 <b>Outcome-based Contract</b>	Cell 4 <b>Behavior-based Contract/ Outcome-based Contract</b>

Figure 1 describes four types of contract arrangements based on contracted services’ task programmability and outcome measurability. In Cell 2, the means-ends relationships involved in agent services can be explicitly observed and specified. As such, information asymmetry between the principal and the agent is low and thus the principal could use behavior-based contracts to purchase agent behaviors. If agent services are ambiguous to observe but their outcomes could be clearly measured with less difficulty (Cell 3), the principal would prefer outcome-based contracts to align the agent’s incentives with those of the principal. When both cause/effect relationships and outcomes are highly certain (Cell 4), both contract types work equally well. The most problematic situation for contract design occurs when agent services share both low task programmability and low outcome measurability (Cell 1). In this circumstance, the focus of principal control seems obscure, making contract design challenging. This is unfortunately where human services usually fit in.

Under the traditional FFS method, a behavior-based contract, government aims to directly define service inputs and process that are necessary to achieve promised results. However, it is always difficult to predict initially what services could exactly lead to desired results due to ambiguous jobs, limited knowledge about the causality, and uncertain future contingencies. Human service provision is highly labor intensive, requiring the exercise of discretionary judgments by service providers (Hasenfeld 1983; Lipsky 1980). Thus, government effort on task programmability under FFS would be offset by the uncertain nature of the service delivery process. And the negotiation nature of human service contracting (DeHoog 1990) would further justify contractors' exercise of discretion. In addition, as the link between the means and the ends of service delivery is broken, contract compensation becomes somewhat independent of service outcomes. If so, contractors have no incentive to improve service performance, because better performance means economic inefficiency (Wulczyn 2005). For example, improving service quality increases contractor costs for advanced facilities and staff, and better services reduce client demands for future services.

The new PBC approach, recognizing the provider discretion in serving clients, evaluates contractors based on service results. In this way, PBC hopes to encourage quality services and better outcomes. But these benefits are subject to two key assumptions—PBC is not vulnerable to (1) measurement problems and (2) gaming by contractors (Behn and Kant 1999; Bevan and Hood 2006). However, human service performance is always ambiguous to capture. First, human services aim to “protect, maintain, or enhance the personal well-being of individuals by defining, shaping, or altering their personal attributes” (Hasenfeld 1983, 1). But beyond such interventions, there are a number of uncontrollable factors out of service providers' reach that would lower the certainty of desired outcomes (DeHoog and Salamon 2002; Martin and Kettner 1996). Second,

human services often pursue multi-dimensional and competing values, such as efficiency, equity, and responsiveness. As such, figuring out appropriate measures that could cover the full spectrum of performance can be difficult (Heinrich and Fournier 2004; Wilson 1989). Third, most human services seek to promote long-term stability and welfare, but performance measures in service contracts have to emphasize short-term effects because all contracts have certain durations and long-term effects are too costly to track. As a result, public managers have to use intermediate outcomes to account for final ones (Heckman et al. 2002; Martin and Kettner 1996). All these imply that performance measures in human service contracting are often just approximations of the targeted outcomes.

Baker (1992; 2002) shows when the principal's objective is not "contractible," the incentives associated with surrogate performance measures are nonoptimal. The more distortion in performance measures, the lower the incentive for desired objectives. This distortion becomes even more severe when gaming enters the picture. As noted above, service contractors embrace discretion when delivering services and PBC even enhances such discretion. Thus, it is very likely that contractors use their information advantage to perversely adjust inaccurate performance measures in order to appear to be behaving well (Courty and Marschke 2004; Moynihan 2011). Williamson (1985, 47) terms this as opportunism, a "self-interest seeking with guile." Bevan and Hood (2006) summarize three forms of gaming problem—ratchet effects (restricting current output to gain undemanding future performance target), threshold effects (downgrading the output of those performing better than the target to meet the target), and output distortions (achieving targeted performance measures at the expense of unmeasured performance). Such performance paradox (Van Thiel and Leeuw 2002) would limit the effectiveness of PBC in human services. By and large, neither a behavior-based nor an outcome-

based contract fits seamlessly with human service provision. The problem here becomes which one is less risky. With this question in mind, I turn to the case study of Indiana's transition from FFS to PBC in the purchase of VR placement services.

### **III. PBC in Vocational Rehabilitation Program**

Vocational Rehabilitation (VR) is a federal-state program that helps individuals with physical or mental disabilities to prepare for, gain, and maintain employment. Title I of the Rehabilitation Act of 1973 authorizes the VR program to empower individuals with disabilities to maximize employment, economic self-sufficiency, independence, and inclusion and integration into society. The assumption is that employment would help disabled people move toward desired quality-of-life changes. The Department of Education provides Title I grants to the state VR agencies to provide employment-related services for individuals with disabilities. This research specifically focuses on job-placement related services in VR, including job search assistance, job placement assistance, and on-the-job support. Often, state VR agencies, through service contracts, acquire these services from community rehabilitation programs.

Traditionally, these contracts are process-oriented, in which government specifies defined services, a purchasable unit for each service (mostly an hour), and a unit cost for each defined service (Revell, West and Cheng 1998). Once a service is delivered, contractors are paid by the amount of services incurred in the service process (Wehman, Revell and Kregel 1998). For example, a contractor may be paid \$30 for each hour of job search service it provides to an eligible service recipient. The rationale behind this hour-based FFS contract design is to customize service needs of people with disabilities and reimburse contractors for providing individualized services. Through intensive reporting by service providers throughout service delivery process, funding agencies try to control the services needed for successful employment

and the detailed flow of funds. However, as noted earlier, the weakness of this FFS design is visible: contractors are compensated for hours of services and thus do not need to consider the results of those services (Novak et al. 1999). Indeed, it is in the contractors' fiscal interest to emphasize service provision and hours billed rather than working toward employment and long-term stability. This further leads to high service costs and poor employment outcomes.

Therefore, there was an incentive for a more effective contracting approach that simultaneously considers valued employment outcomes and the costs to achieve those outcomes. In 1990s, the Oklahoma Department of Rehabilitation Services first launched the PBC model, or what they call Milestone Contracting System, in the purchase of placement services (Frumkin 2001). Under its PBC system, service contractors are reimbursed at a fixed amount when clients they serve successfully achieve a sequential series of intermediate performance milestones – establish job goal, become employed, stabilize in employment, and continue in employment. State agencies do not specify vocational methods and amount of services to be used. Contractors have flexibility and incentives in achieving specified milestones rapidly. In this way, PBC is expected to generate a triple win for VR programs: disabled people receiving quick and quality services, contractors enjoying less regulation and greater flexibility, and state VR agencies achieving better results at lower costs with greater accountability (Frumkin 2001; O'Brien and Revell 2005). The Oklahoma model soon received extensive recognition and replication by other state VR programs (O'Brien & Revell, 2005). Among those states using PBC, Indiana is the latest one in the transition, which provides a case that allows using administrative data to examine PBC effectiveness.

The Indiana Bureau of Rehabilitation Services changed its statewide contracting approach for placement services from FFS into PBC, which they refer to as result-based funding,

in late 2006. Before the statewide initiation, a pilot project was undertaken to engage stakeholders in framing the design of the RBF system, including milestone setting, criteria to verify milestone attainment, reimbursement rates for each milestone, etc. (McGrew et al. 2005). The emphasis of RBF was placed upon structuring service contracting method that would increase the likelihood of both initial job placement and long-term tenure. Under RBF, contractors receive reimbursement at a fixed rate once consumers reach predetermined stages across the employment process, with the higher payments toward the later milestones. The rate was determined by the government agency, factoring provider-estimated cost at each milestone. The total amount paid for all the milestones reflected the statewide historical average costs in the purchase of each successful case closure under the traditional FFS approach, plus an amount equal to the average service costs of those that failed to help clients reach case closure. In order to address the client selection problem, the system includes two tiers of payment rates, tailored to clients with different degrees of disability. The VR counselor makes the decision on milestone authorization and the tier the individual will enter. Each milestone will be pre-authorized by the counselor and will be paid only once per case, per contractor, upon receipt and acceptance of the required documentation for payment by the counselor. Figure 2 shows an example of the Indiana RBF system.

**Figure 2. A Demonstration of Indiana Result-based Funding System**

<b>Milestone</b>	<b>Tier I Rate</b>	<b>Tier II Rate</b>	<b>Outcome Description</b>
1. Plan for employment & supports	\$1,200	\$ 600	A Plan for employment and supports developed by the customer and his/her support team (including the customer, VR counselor, service provider, and any other stakeholders).
2. Job placement	\$1,200	\$ 900	The customer has worked one week at the hours per weekly work goal in the vocational area identified in the Plan.

3. Four-week placement	\$1,864	\$1,325	The customer has worked four weeks in which he/she met hours per weekly work goal and pay rate as stated in the Plan. The customer and the employer have indicated satisfaction.
4. Eligible for closure	\$4,000	\$2,600	The customer has maintained employment for 60 calendar days for Supported Employment or 90 calendar days for others. The customer is employed in a job as outlined in his/her Plan that is commensurate with his/her skills and abilities. Customer and employer are satisfied.
<b>TOTAL</b>	<b>\$8,264</b>	<b>\$5,425</b>	

Note: Tier I is used for people who (1) qualifies as the most severely disabled as defined in the state policy, (2) requires multiple services over an extended period of time, and (3) is likely to need ongoing, intensive intervention to get and keep a job. Tier II is for people who (1) has a disability, severe disability, or most severe disability, and (2) would not require ongoing, intensive intervention to get and keep a job.

Source: Indiana Bureau of Rehabilitation Services.

Although the performance of a contractual network is a multi-dimensional construct (Provan and Milward 2002), this paper evaluates the effectiveness of PBC only from a service outcome perspective, i.e., if PBC contributes to the improvement in client well-being. The unit of analysis is the individual client receiving placement services. Several approximations of employment outcomes are identified: likelihood of getting employed, time to placement, job retention, and wage. The first two are directly targeted by the performance measurement. PBC motivates contractors to move across the performance milestones quickly in order to receive reimbursement. Thus, I predict:

H<sub>1</sub> After using PBC, clients are more likely to attain employment.

H<sub>2</sub> After using PBC, clients are able to achieve employment in less time.

In addition to these two indicators, another two employment quality indicators (job retention and wage) are also included to examine if the potential performance improvement in employment possibility and time-to-placement is attained through gaming other unmeasured performance. In the above PBC system, employment quality is not directly targeted, only implied as a threshold, such as state minimum hourly wage. By giving high discretion to contractors during the service process, PBC assumes contractors would work with clients meticulously and innovatively and help them secure high-quality employment. Thus, this research would also test:

H<sub>3</sub> After using PBC, clients are able to achieve longer job retention.

H<sub>4</sub> After using PBC, clients are able to gain higher wages.

#### **IV. Research Design**

This paper uses a quasi-experimental design (Figure 3) to evaluate the treatment effect of PBC. It compares individual employment outcomes in the Indiana VR program before and after the PBC intervention within the time period 2004-2009, with Michigan VR program as a control. As mentioned, Indiana, as the treatment group, changed the funding mechanism in the purchase placement services from FFS to PBC at the end of 2006. Michigan, Indiana's neighbor state, still used the traditional FFS method during that time period, reimbursing contractors based on the amount of services incurred. The repeated cross-sectional data was requested from the Rehabilitation Services Administration (RSA) of the Department of Education. The RSA 911 database reports records pertaining to all the individuals whose case records were closed in a given fiscal year, including personal characteristics, types of services, and employment outcome of all clients receiving state VR services.

**Figure 3. Interrupted time series with a nonequivalent control group design**

	2004	2005	2006		2007	2008	2009
IN	O <sub>1</sub>	O <sub>2</sub>	O <sub>3</sub>	X	O <sub>4</sub>	O <sub>5</sub>	O <sub>6</sub>
MI	O <sub>1</sub>	O <sub>2</sub>	O <sub>3</sub>		O <sub>4</sub>	O <sub>5</sub>	O <sub>6</sub>

Note: O<sub>n</sub> represents the observation at time point n, and X represents the treatment.

Intuitively, the treatment effect is the net difference between the condition of a unit after receiving a treatment and the condition of that unit if it would have not received that treatment. However, these two conditions are not possible to observe at the same time, which constitutes what Holland (1986, 947) calls the “fundamental problem of casual inference.” The core task of policy evaluation, thus, turns to the construction of a counterfactual outcome to estimate the unobserved outcome. This further implies forming a control group that to a greatest extent approximates the treatment one in various aspects. Admittedly, randomized experimentation is most robust in this regard. By assigning subjects randomly to experimental conditions, experimenters could guarantee that there are no systematic differences on all background covariates between comparison groups before treatment. However, for various ethical and practical reasons, this ideal randomized experimentation is not feasible in current study. Thus, quasi-experimentation was adopted, in which samples are collected through observation (Shadish, Cook and Campbell 2002). As such, it is very likely that “the treated and control groups differ prior to treatment in ways that matter for the outcomes under study” (Rosenbaum 2002, 71). In this sense, the systematic pre-treatment differences may bias the internal validity of causal inference. Here, I rely on Campbell and Stanley’s (1963) typology on the threats to internal validity in quasi-experimental designs.

The interrupted time series with a nonequivalent control group design used in this study is robust in removing most of the threats to internal validity, such as maturation, testing, and

regression, but still retaining instrumentation, selection bias, and local context (Shadish et al. 2002). Thus, these three potential threats should be minimized as much as possible before comparing two groups. Instrumentation may bias causal inference when different administrative procedures and measures are used to record participants' performance over time. However, this would not be a big concern for state VR programs. Under the Rehabilitation Act, all the administrative and service components, procedures, and standards are under rigorous federal regulations. For example, RSA conducts annual reviews and periodic on-site monitoring of state VR programs to ensure they comply with program and performance requirements under the Rehabilitation Act. In this way, the consistency within and between states can be expected.

A serious threat comes from selection bias, i.e., differences exist between individuals in treatment and control groups. To solve this problem, matched sampling is used to correct the observed imbalances between the two states. Matched sampling is a resampling strategy, “selecting units from a large reservoir of potential controls to produce a control group of modest size that is similar to a treated group with respect to the distribution of observed covariates” (Rosenbaum and Rubin 1985, 33). After matching, two comparison groups are identical on a variety of observed variables, which actually replicates a randomized experiment where the treatment assignment is unconfounded, at least given the observed covariates (Rosenbaum and Rubin 1983; Rubin 1973). In particular, this study adopts propensity score matching to produce the matched sample. A propensity score, as Rosenbaum and Rubin (1983, 41) define, is “the conditional probability of assignment to a particular treatment given a vector of observed covariates.” Matching samples based on propensity scores allows simultaneously considering a variety of covariates. Rather than requiring exact or close matching on all covariates separately, propensity score matching enables matching on the scalar summary of the covariates. Given a

propensity score, the differences in the observed covariates between a treatment unit and a control unit are balanced.<sup>3</sup> Therefore, matching treatment and control units with the same propensity scores could create new comparison groups with identical distribution of observed covariates (Rosenbaum 2002).

Local context might also bias causal inference when the individuals in comparison groups reside in different settings. To address this issue, this study chooses Michigan as the control group against Indiana, aiming to maximize the socio-economic similarities between the two states. In addition, I use difference-in-differences (DID) models after matched sampling to further adjust the unobserved imbalance. Under the DID model, any bias caused by exogenous variables common to Indiana and Michigan could implicitly be controlled for, even when these variables are unobserved.<sup>4</sup> Indeed, running DID regressions on matched samples embraces a number of advantages. First, the combination of the two methods is most robust and efficient in removing the biases due to covariates and estimating the treatment effect on the treated (Abadie and Imbens 2006; Heckman, Ichimura and Todd 1997; Rubin 1973, 1979). Matched sampling substantially reduces observed covariate differences, and model-based adjustment after further controls for residual differences. Second, matched sampling relaxes the DID identification restrictions, making model-based adjustment less sensitive to model specification. This again allows the estimation of parsimonious parametric approximations of the average treatment effect on the treated. (Abadie 2005; Ho, Imai, King and Stuart 2007).

---

<sup>3</sup> It is possible that two units with the same propensity score may be different in a certain observed covariate, but those differences are not systematic (Guo and Fraser 2010).

<sup>4</sup> The common trend assumption is a very strong one, but I do have some evidence to support the assumption during 2004-2009. I checked the state-level factors that might affect employment outcomes, including GDP growth, unemployment rate, average weekly earnings, and VR program capacity (measured by average number of clients served per program staff), and found they roughly follow the similar trend. I also reviewed the annual review reports of Indiana and Michigan VR programs and didn't find major policy changes on the purchase of employment services.

## V. Data Analysis and Results

Propensity score matching was first used to produce matched samples, following the procedures suggested by Caliendo and Kopeinig (2008) and Guo and Fraser (2010). To begin with, three categories of covariates in Table 1—demographic background (age, education, race, gender, veteran status, primary disability, and secondary disability), pre-service status (employment status, work disincentives, previous service status, and Projects with Industry status), and employment service (number of placement related services received)—that are theoretically relevant to employment outcomes were included into the conditioning model to estimate propensity scores for individual clients in the two states. When selecting covariates, I followed Rubin and Thomas (1996, 253) that “unless a variable can be excluded because there is a consensus that it is unrelated to outcome or is not a proper covariate, it is advisable to include it in the propensity score model even if it is not statistically significant.”

[Table 1 Here]

Propensity scores were then estimated using a binary logistic regression with these three groups of covariates as independent variables and treatment assignment (Indiana=1, Michigan=0) as dependent variable<sup>5</sup>. After that, treatment individuals were matched with control individuals based on the values of propensity scores, using 1-to-1 nearest neighbor matching without replacement. A caliper, a quarter of a standard deviation of the propensity scores of the sample (Rosenbaum and Rubin 1985), was added to ensure that matched units were chosen only when

---

<sup>5</sup> I used STATA program `pscore.ado` developed by Becker and Ichino (2002) to estimate propensity scores. In particular, this program helps ensure the balancing property of propensity scores, i.e. observations with the same propensity scores should have the same distribution of observed characteristics, regardless of treatment status.

the absolute distance between the two units was within the caliper. After matching, it is expected that the preexisted observed differences between the two states are substantially eliminated. Before moving forward, I checked the covariate balance using the absolute standardized difference of covariate means and t-tests. Appendix I shows the results of covariate balance check. Ideally, after matching the absolute standardized difference of covariate means should be less than 5% and t-statistic should no longer be significant (D'Agostino 1998; Haviland, Nagin and Rosenbaum, 2007). In this vein, the matched sampling in this study is quite effective in removing a substantial portion of the preexisting differences between the two states, but not all of them, as expected.<sup>6</sup>

With the matched sample, DID models were run to estimate the impact of PBC on Indiana clients. The DID models are specified as follows:

For logistic model on employment probability:

$$P(Y = 1) = F[\beta_0 + \beta_1(\text{Indiana}) + \beta_2(\text{Service Year } 2007 - 2009) + \beta_3(\text{Indiana} * \text{Service Year } 2007 - 2009) + \beta_4X_1 + \beta_5X_2 + \beta_6X_3]$$

For ordinary least squares (OLS) models on time to placement, weekly working hours, and weekly earnings:

$$Y = \beta_0 + \beta_1(\text{Indiana}) + \beta_2(\text{Service Year } 2007 - 2009) + \beta_3(\text{Indiana} * \text{Service Year } 2007 - 2009) + \beta_4X_1 + \beta_5X_2 + \beta_6X_3 + \mu$$

$X_1$  contains “demographic background” variables, including age, education, race, gender, primary disability, and secondary disability.

---

<sup>6</sup> A major problem in matched sampling is inexact matching: it is not always possible to find enough matched treatment and control samples with exactly the same observed covariates, especially when the number of covariates increases (Rubin 1979).

X<sub>2</sub> contains “pre-service status” variables, including employment status, work disincentives, previous service status, and participation in projects with Industry.

X<sub>3</sub> contains “employment services” variable, i.e., number of placement services received.

[Tables 2.1 and 2.2 Here]

Tables 2.1 and 2.2 present DID regression results. Within each model, the interaction term ( $\beta_3$ ) between the variable of *Indiana* and the variable of *service period 2007-2009* is the DID estimator of the treatment effect on the treated. First, logistic regression was employed to predict the differences in the likelihood of attaining employment results for those who received employment services before and after PBC. Tests of goodness of fit of the regression model were also performed. Overall, the logistic regression model is statistically significant (likelihood ratio chi-square=1102.74, p= .0000; Hosmer-Lemeshow chi-square = 8.943, p= .063), showing the model is reliable to produce meaningful inference. Generally, after the introduction of PBC, Indiana clients experienced much higher employment possibilities (odds ratio=1.4991, p< .01)<sup>7</sup>. This finding supports the hypothesis that PBC is significantly more effective in promoting employment results.

Second, OLS regressions were run to compare three performance indicators of employment outcomes before and after PBC: (1) time to placement, (2) weekly working hours, and (3) weekly earnings (adjusted by inflation). Before regression analyses, a series of regression

---

<sup>7</sup> The interpretation of interaction effects in nonlinear models is still under econometric discussion (Ai and Norton 2003; Athey and Imbens 2006; Greene 2010). Ai and Norton (2003) argued that in nonlinear models the marginal effect of the interaction term does not represent the magnitude of the interaction effect. The interaction effect depends on all the covariates, and thus requires computing the cross derivative of the expected value of the dependent variable. The statistical significance of the interaction effect should be based on the estimated cross-partial derivative. Puhani (2012) and Karaca-Mandic, Norton, and Dowd (2012) further demonstrated that under DID context only, the incremental effect of the coefficient of the interaction term could approximate the treatment effect on the treated. I followed this suggestion in this paper.

diagnostics were conducted to ensure the basic assumptions of OLS regression were met. Particularly, robust standard errors were used to correct for the heteroscedasticity of the residuals. Overall, each model is statistically significant, explaining a substantial portion of the variations of the dependent variable respectively. The regression model on time to placement shows individual employees in Indiana after the use of PBC spent 72 days ( $p < .01$ ) less to achieve employment outcomes, which is consistent with the hypothesis that PBC motivates service contractors to achieve employment outcomes rapidly. The models on employment quality (working hours and wages) demonstrate mixed results. The same as the hypothesis on job retention, Indiana clients under PBC worked 1.33 hours ( $p < .05$ ) longer than their counterparts weekly during 2004-2006. The hypothesis on wages is partially supported. Weekly wages of Indiana employees increased by \$4.37 after the introduction of PBC, but the difference is not statistically significant even at  $p < .1$  level. However, these small differences in working hours and wages, though meaningful in statistical sense, are actually of little real policy significance.

## **VI. Conclusion**

The managerial motivation behind all performance-based strategies is the phrase that “what gets measured gets done.” The introduction of PBC in human service provision is also the case. By attaching contract compensation to performance achievements, PBC urges contractors to achieve desired outcomes in a timely manner. Because of the outcome orientation, PBC gives service contractors considerable discretion throughout the service process, expecting them to use innovative and quality customer services to further enhance service outcomes. This study tests these two claims by comparing the employment outcomes under PBC and FFS through a quasi-experiment. As predicted, PBC performs much better in measured performance, demonstrating

higher chances of employment and shorter time-to-placement. But the differences in unmeasured performance (working hours and wages) are trivial. It seems that PBC is better than FFS in that it achieves desired employment milestones in a more efficient way, without degrading employment quality.

The implication of this study is worth mentioning. The research project shows the dynamics and the difficulty of introducing performance management to human service contracting. Largely due to human services' ambiguous performance and high provider discretion, the use of PBC in human service contracting should be very careful. The finding here may further raise the policy concern on how to take full advantage of PBC, particularly when we have found it is somewhat better than FFS. Technically, there are a number of ways to improve the PBC design, such as fixing performance measures and changing incentive structures. For example, Hill's (2006) study of casework task configurations in welfare-to-work programs proposed that the separation of measurable and unmeasurable tasks among frontline workers would contribute to program effectiveness. Heinrich and Choi (2007) suggested changing performance measures periodically before contractors learn the ways to gaming the measures. However, these technical efforts on optimizing PBC systems would hardly be free themselves from the puzzle of human service performance mentioned previously. Another way to improve PBC might be a managerial one.

The research and the practice of PBC tend to ignore the two faces of contracting. The formal side of contracting, such as PBC, considers designing formal arrangements to structure contracting behaviors and performance. In addition, the informal side of contracting fulfills contractual agreements through relational sanction and social interaction. It highlights informal relationship building and trust cultivation between contractual parties (Macaulay, 1963; Poppo

and Zenger 2002; Van Slyke 2007). As mentioned earlier, the design of formal contract arrangement for human services is demanding. As such, social control, relying on informal and relational mechanisms, may emerge to function as a supplement to align the interests between the principal and the agent and encourage appropriate behaviors. Generally in human services, not all aspects of performance can be clearly defined and measured. In this way, the use of PBC with surrogate measures as a formal mechanism might inevitably lead to incomplete performance improvement or even gaming. More broadly, the more provider discretion and ambiguous service performance involved in human service delivery, the less effectiveness PBC as a formal control mechanism could produce, which leaves room for relational contracting to fit in. In short, public managers should pay attention to the relational side of contracting and devote administrative resources to building relationships and trust with contractors to support PBC endeavor.

This study here may suffer from several limitations. First, although I used propensity score matching and DID models jointly to balance the two states during the time period of interest, I still cannot guarantee the exact similarity in the two states over time. This is particularly a concern when using quasi-experiments in cross-state comparison (Heckman et al. 1997; Michalopoulos, Bloom, and Hill 2004). Second, due to data constraint, I failed to address three important indicators of PBC effectiveness in placement services. Client selection problem was not addressed in that the database here only records the individuals who had been already admitted into service process. The costs to achieve employment outcomes (including service costs in the purchase of contracted services and administrative costs in monitoring contractors) under two contracting models were not compared, either. Also, the long-term employment effect was not examined, because in the database successful case closures happen mostly when clients

attain 90-day employment. Third, the external validity of this project as a case study, as Yin (2009) suggests, lies in “analytical generalization” through replication in different contexts rather than “statistical generalization” through inference from a sample to a population. In this sense, the findings here might be used only for conditional, contingent generalizations (George and Bennett 2005) to other cases which are similar to the one under study. The robustness of the conclusion awaits other future studies to triangulate.

## References:

- Abadie, A. 2005. Semiparametric difference-in-differences estimators. *Review of Economic Studies* 72: 1-19.
- Abadie, A., and G. W. Imbens. 2006. Large sample properties of matching estimators for average treatment effects. *Econometrica* 74: 235-67.
- Ai, C., and E. C. Norton. 2003. Interaction terms in logit and probit models. *Economics letters* 80: 123-29.
- Amirkhanyan, A. A. 2010. Monitoring across sectors: Examining the effect of nonprofit and for-profit contractor ownership on performance monitoring in state and local contracts. *Public Administration Review* 70: 742-55.
- Athey, S., and G. W. Imbens. 2006. Identification and inference in nonlinear difference-in-differences models. *Econometrica* 74: 431-97.
- Arrow, K. J. 1964. Control in large organizations. *Management Science* 10: 397-408.
- Baker, G. P. 1992. Incentive contracts and performance measurement. *Journal of Political Economy* 100: 598-614.
- Baker, G. P. 2002. Distortion and risk in optimal incentive contracts. *Journal of Human Resources* 37: 728-51.
- Barnow, B. S. 2000. Exploring the relationship between performance management and program impact: A case study of the Job Training Partnership Act. *Journal of Policy Analysis and Management* 19: 118-41.
- Becker, S. O., and A. Ichino. 2002. Estimation of average treatment effects based on propensity scores. *Stata Journal* 2: 358-77.
- Behn, R. D. 2003. Why measure performance? Different purposes require different measures. *Public Administration Review* 63: 586-606.
- Behn, R. D., and P. A. Kant. 1999. Strategies for avoiding the pitfalls of performance contracting. *Public Productivity & Management Review* 22: 470-89.
- Bevan, G., and C. Hood. 2006. What's measured is what matters: Targets and gaming in the English public health care system. *Public Administration* 84: 517-38.
- Brown, T. L., and M. Potoski. 2004. Managing the public service market. *Public Administration Review* 64: 656-68.
- Brucker, D. L., and M. Stewart. 2011. Performance-based contracting within a state substance abuse treatment system: A preliminary exploration of differences in client access and client outcomes. *Journal of Behavioral Health Services & Research* 38: 383-97.
- Caliendo, M., and S. Kopeinig. 2008. Some practical guidance for the implementation of propensity score matching. *Journal of Economic Surveys* 22: 31-72.
- Campbell, D. T., and J. C. Stanley. 1963. *Experimental and quasi-experimental designs for research*. Boston: Houghton Mifflin.
- Commons, M., T. G. McGuire, and M. H. Riordan. 1997. Performance contracting for substance abuse treatment. *Health Services Research* 32: 631-50.
- Cooper, P. J. 2003. *Governing by contract: Challenges and opportunities for public managers*. Washington, DC: CQ Press.
- Courty, P., and G. Marschke. 2004. An empirical investigation of gaming responses to explicit performance incentives. *Journal of Labor Economics* 22: 23-56.
- D'Agostino, R. B., Jr. 1998. Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Statistics in Medicine* 17: 2265-81.

- DeHoog, R. H. 1990. Competition, negotiation, or cooperation. *Administration & Society* 22: 317-40.
- DeHoog, R. H., and L. M. Salamon. 2002. Purchase-of-service contracting, 319-39. In *The tools of government: A guide to the new governance*, ed. L. M. Salamon. New York: Oxford University Press.
- Dias, J. J., and S. Maynard-Moody. 2007. For-profit welfare: Contracts, conflicts, and the performance paradox. *Journal of Public Administration Research and Theory* 17: 189-211.
- Eisenhardt, K. M. 1989. Agency theory: An assessment and review. *Academy of Management Review* 14: 57-74.
- Frederickson, D. G., and H. G. Frederickson. 2006. *Measuring the performance of the hollow state*. Washington, D.C.: Georgetown University Press.
- Frumkin, P. 2001. *Managing outcomes: Milestone contracting in Oklahoma*. Washington, DC: The IBM Center for The Business of Government.
- George, A. L., and A. Bennett. 2005. *Case studies and theory development in the social sciences*. Cambridge, MA: MIT Press.
- Greene, W. 2010. Testing hypotheses about interaction terms in nonlinear models. *Economics Letters* 107: 291-96.
- Guo, S., and M. W. Fraser. 2010. *Propensity score analysis: Statistical methods and applications*. Thousand Oaks, CA: Sage.
- Hasenfeld, Y. 1983. *Human service organizations*. Englewood Cliffs, NJ: Prentice-Hall.
- Hatry, H. P., and J. S. Wholey. 1999. *Performance measurement: Getting results*. Washington, DC: Urban Institute Press.
- Haviland, A., D. S. Nagin, and P. R. Rosenbaum. 2007. Combining propensity score matching and group-based trajectory analysis in an observational study. *Psychological Methods* 12: 247-67.
- Heckman, J., C. Heinrich, and J. Smith. 2002. The performance of performance standards. *Journal of Human Resources* 37: 778-811.
- Heckman, J. J., H. Ichimura, and P. E. Todd. 1997. Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme. *The Review of Economic Studies* 64: 605-54.
- Hefetz, A., and M. Warner. 2004. Privatization and its reverse: Explaining the dynamics of the government contracting process. *Journal of Public Administration Research and Theory* 14: 171-90.
- Heinrich, C. J., and Y. Choi. 2007. Performance-based contracting in social welfare programs. *American Review of Public Administration* 37: 409-35.
- Heinrich, C. J., and E. Fournier. 2004. Dimensions of publicness and performance in substance abuse treatment organizations. *Journal of Policy Analysis and Management* 23: 49-70.
- Hill, C. J. 2006. Casework job design and client outcomes in welfare-to-work offices. *Journal of Public Administration Research and Theory* 16: 263-88.
- Ho, D., K. Imai, G. King, and E. Stuart. 2007. Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis* 15: 199-236.
- Holland, P. W. 1986. Statistics and causal inference. *Journal of the American Statistical Association* 81: 945-60.

- Jensen, M. C., and W. H. Meckling. 1976. Theory of the firm: Managerial behavior, agency costs and ownership structure. *Journal of Financial Economics* 3: 305-60.
- Johnston, J. M., and B. S. Romzek. 1999. Contracting and accountability in state medicaid reform: Rhetoric, theories, and reality. *Public Administration Review* 59: 383-99.
- Karaca-Mandic, P., E. C. Norton, and B. Dowd. 2012. Interaction terms in nonlinear models. *Health Services Research* 47: 255-74.
- Kerr, S. 1975. On the folly of rewarding A, while hoping for B. *Academy of Management Journal* 18: 769-83.
- Kettl, D. F. 1993. *Sharing power: Public governance and private markets*. Washington, DC: The Brookings Institution.
- Kettl, D. F. 2002. Managing indirect government, 490-510. In *The tools of government: A guide to the new governance*, ed. L. M. Salamon. New York: Oxford University Press.
- Kravchuk, R. S., and R. W. Schack. 1996. Designing effective performance-measurement systems under the Government Performance and Results Act of 1993. *Public Administration Review* 56: 348-58.
- Lambright, K. T. 2009. Agency theory and beyond: Contracted providers' motivations to properly use service monitoring tools. *Journal of Public Administration Research and Theory* 19: 207-27.
- Lipsky, M. 1980. *Street-level bureaucracy: Dilemmas of the individual in public services*. New York: Russell Sage.
- Lu, M. 1999. Separating the true effect from gaming in incentive-based contracts in health care. *Journal of Economics and Management Strategy* 8: 383-431.
- Macaulay, S. 1963. Non-contractual relations in business: A preliminary study. *American Sociological Review* 28: 55-67.
- Martin, L. L., and P. M. Kettner. 1996. *Measuring the performance of human service programs*. Thousand Oaks, CA: Sage.
- Martin, L. L. 2005. Performance-based contracting for human services: Does it work? *Administration in Social Work* 29: 63-77.
- McBeath, B., and W. Meezan. 2010. Governance in motion: Service provision and child welfare outcomes in a performance-based, managed care contracting environment. *Journal of Public Administration Research and Theory* 20: i101-23.
- McGrew, J. H., J. K. Johannesen, M. E. Griss, D. L. Born, and C. Katuin. 2005. Performance-based funding of supported employment: A multi-site controlled trial. *Journal of Vocational Rehabilitation* 23: 81-99.
- McLellan, A. T., J. Kemp, A. Brooks, and D. Carise. 2008. Improving public addiction treatment through performance contracting: The Delaware experiment. *Health Policy* 87: 296-308.
- Michalopoulos, C., H. S. Bloom, and C. J. Hill. 2004. Can propensity-score methods match the findings from a random assignment evaluation of mandatory welfare-to-work programs? *Review of Economics and Statistics* 86: 156-79.
- Milward, H. B., and K. G. Provan. 2000. Governing the hollow state. *Journal of Public Administration Research and Theory* 10: 359-80.
- Moynihan, D. P. 2011. The promises and paradoxes of performance-based bureaucracy, 278-302. In *The Oxford handbook of American bureaucracy*, ed. R. F. Durant. New York: Oxford University Press.
- Novak, J., D. Mank, G. Revell, and N. Zemaitis. 1999. Initiatives influencing the emergence of results-based funding of supported employment services, 25-42. In *The impact of*

- supported employment for people with significant disabilities*, ed. G. Revell, K. J. Inge, D. Mank, and P. Wehman. Richmond, VA: Rehabilitation Research & Training Center on Workplace Supports.
- O'Brien, D., and G. Revell. 2005. The milestone payment system: Results based funding in vocational rehabilitation-2005. *Journal of Vocational Rehabilitation* 23: 101-14.
- Ouchi, W. G. 1979. A conceptual framework for the design of organizational control mechanisms. *Management Science* 25: 833-48.
- Ouchi, W. G., and M. A. Maguire. 1975. Organizational control: Two functions. *Administrative Science Quarterly* 20: 559-69.
- Poppo, L., and T. Zenger. 2002. Do formal contracts and relational governance function as substitutes or complements? *Strategic Management Journal* 23: 707-25.
- Provan, K. G., and H. B. Milward. 2002. Do networks really work? A framework for evaluating public-sector organizational networks. *Public Administration Review* 61: 414-23.
- Puhani, P. A. 2012. The treatment effect, the cross difference, and the interaction term in nonlinear "difference-in-differences" models. *Economics Letters* 115: 85-7.
- Revell, W. G., M. West, and Y. Cheng. 1998. Funding supported employment. *Journal of Disability Policy Studies* 9: 59-79.
- Romzek, B. S., and J. M. Johnston. 2005. State social services contracting: Exploring the determinants of effective contract accountability. *Public Administration Review* 65: 436-49.
- Rosenbaum, P. R. 2002. *Observational studies*. New York: Springer.
- Rosenbaum, P. R., and D. B. Rubin. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika* 70: 41-55.
- Rosenbaum, P. R., and D. B. Rubin. 1985. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *American Statistician* 39: 33-8.
- Rubin, D. B. 1973. Matching to remove bias in observational studies. *Biometrics* 29: 159-83.
- Rubin, D. B. 1979. Using multivariate matched sampling and regression adjustment to control bias in observational studies. *Journal of the American Statistical Association* 74: 318-28.
- Rubin, D. B., and N. Thomas. 1996. Matching using estimated propensity scores: Relating theory to practice. *Biometrics* 52: 249-64.
- Salamon, L. M. 2002. *The tools of government: A guide to the new governance*. New York: Oxford University Press.
- Shadish, W. R., T. D. Cook, and D. T. Campbell. 2002. *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin.
- Shen, Y. 2003. Selection incentives in a performance-based contracting system. *Health Services Research* 38: 535-52.
- Smith, S. R., and M. Lipsky. 1993. *Nonprofits for hire: The welfare state in the age of contracting*. Cambridge, MA: Harvard University Press.
- Thompson, J. D. 1967. *Organizations in action: Social science bases of administrative theory*. New York: McGraw-Hill.
- Van Slyke, D. M. 2003. The mythology of privatization in contracting for social services. *Public Administration Review* 63: 296-315.
- Van Slyke, D. M. 2007. Agents or stewards: Using theory to understand the government-nonprofit social service contracting relationship. *Journal of Public Administration Research and Theory* 17: 157-87.

- van Thiel, S., and F. L. Leeuw. 2002. The performance paradox in the public sector. *Public Performance & Management Review* 25: 267-81.
- Wehman, P., G. Revell, and J. Kregel. 1998. Supported employment: A decade of rapid growth and impact. *American Rehabilitation* 24: 31-43.
- Williamson, O. E. 1985. *The economic institutions of capitalism*. New York: Free Press.
- Wilson, J. Q. 1989. *Bureaucracy: What government agencies do and why they do it*. New York: Basic Books.
- Wulczyn, F. 2005. *Performance-based contracting: The basics*. Hackensack, NJ: Foster Family-based Treatment Association.
- Yin, R. K. 2009. *Case study research: Design and methods*. Los Angeles, CA: Sage.

**Table 1. Description of Variables**

<b>Variables</b>	<b>Operationalization and Measurement</b>
<b>DEPENDENT VARIABLES</b>	
Employment outcome	An individual's employment status after receiving services, with 0=without employment outcome after services, 1=with employment outcome after services
Time to placement	The number of days an individual spent in achieving an employment outcome, measured by date of closure - date of individualized plan for employment + 1
Weekly working hours	The number of hours an individual worked for earnings in a typical week when an employment outcome was achieved
Weekly earnings	The amount of money an individual earned in a typical week after achieving an employment outcome, by dollar amount adjusted by inflation
<b>INDEPENDENT VARIABLES</b>	
<i>State and Service Year</i>	
State (Indiana)	State name, with 0=Michigan, 1=Indiana
Service Year 2007-2009	If an individual case was closed in fiscal years 2007-2009, with 0=no, 1=yes
Indiana Service Year 2007-2009	The interaction between State Indiana and individual case closed in fiscal years 2007-2009
<i>Demographic Background</i>	
Age	An individual's age at service application
Education	An individual's level of education attained at application, with 0=less than high school, 1=special education, 2=high school graduate, 3=post-secondary/associate degree, and 4=college degree or higher
Race	An individual's race and ethnicity, with 0=black or African American, 1=Native American (American Indian, Alaska native, native Hawaiian, or other Pacific Islander), 2=Asian, 3=white, 4=Hispanic or Latino
Gender (Female)	An individual's gender status, with 0=male, 1=female
Veteran	An individual's veteran status, 0=not a veteran, 1=veteran
Primary disability	An individual's primary physical or mental impairment, with 0=sensory/communication impairments, 1=physical impairments, 2=mental impairments
Secondary disability	An individual's second physical or mental impairment, with 0=no impairment, 1=sensory/communication impairments, 2=physical impairments, 3=mental impairments

<b><i>Pre-service Status</i></b>	
Employment status	An individual's employment status at application, with 0=not employed, 1=employment
Work disincentives	The number of public support an individual had at application, including supplemental security income (SSI), Temporary Assistance for Needy Families (TANF), general assistance from state or local government, social security disability insurance (SSDI), veterans' disability benefits, workers' compensation, Medicaid, Medicare, medical insurance not through employment, and others
Previous service status	If an individual had received previous employment service, with 0=no previous closure, 1= closed before services, 2=closed after services
Participation in Projects with Industry	If an individual participates in Projects with Industry program, with 0=no, 1=yes
<b><i>Employment Services</i></b>	
No. of placement services received	The number of employment services an individual received throughout service process, including job search assistance, job placement assistance, and on-the-job supports

**Table 2.1 Logistic Regression Model Predicting Likelihood of Employment for Service Recipients (N = 12, 372)**

<b>Variable</b>	<b>Odds Ratio</b>	<b>Standard Error</b>	<b>Z Value</b>
<i>State and Service Year</i>			
State (Indiana)	0.7561***	0.0397	-5.33
Service Year 2007-2009	0.8287***	0.0450	-3.46
Indiana Service Year 2007-2009	1.4991***	0.1144	5.30
<i>Demographic Background</i>			
Age	0.9975	0.0016	-1.55
Education			
Special education	1.3384***	0.1207	3.23
High school graduate	1.2281**	0.1099	2.30
Post-secondary/associate degree	1.2460**	0.1249	2.19
College degree or higher	1.4369***	0.1783	2.92
Race			
Native American	0.4991**	0.1455	-2.38
Asian	1.4769*	0.3241	1.78
White	1.3715***	0.0707	6.13
Hispanic or Latino	1.4601***	0.1847	2.99
Gender (Female)	0.8292***	0.0326	-4.76
Veteran	0.8319*	0.0846	-1.81
Primary disability			
Physical impairments	0.7305***	0.0810	-2.83
Mental impairments	1.0239	0.1065	0.23
Secondary disability			
Sensory/communication impairments	1.0099	0.01084	0.09
Physical impairments	0.8492***	0.0480	-2.89

Mental impairments	0.8156***	0.0358	-4.46
<b><i>Pre-service Status</i></b>			
Currently employed	2.0271***	0.1071	13.38
Work disincentives	0.9278***	0.0160	-4.36
Previous closure/service			
Closed before services	0.8738*	0.0650	-1.81
Closed after services	1.2990***	0.0640	5.31
Participation in Projects with Industry	3.4439***	1.4556	2.93
<b><i>Employment Services</i></b>			
No. of placement services received	2.8868***	0.1392	21.99
Likelihood ratio chi square	1102.74***		
Pseudo $R^2$	.2653		

\*significant at .1; \*\*significant at .05; \*\*\*significant at .01; two-tailed tests.

**Table 2.2 OLS Regression Models Analyzing Employment Outcomes (N = 4, 940)**

Variable	Model (1) Time to placement			Model (2) Weekly working hours			Model (3) Weekly earnings		
	Coefficient	Robust Standard Error	t-value	Coefficient	Robust Standard Error	t-value	Coefficient	Robust Standard Error	t-value
<i>State and Service Year</i>									
State (Indiana)	123.7927***	9.69203	12.77	-2.7736***	0.3925	-7.07	-40.4802***	5.4046	-7.49
Service Year 2007-2009	28.1780***	9.7877	2.88	-1.2952***	0.3868	-3.35	2.3857	5.4851	0.43
Indiana Service Year 2007-2009	-72.1985***	14.3650	-5.03	1.3291**	0.5603	2.37	4.3715	7.2722	0.60
<i>Demographic Background</i>									
Age	-2.3190***	0.31189	-7.44	-0.0200	0.0122	-1.63	0.4703***	0.1642	2.86
Education									
Special education	-0.5269	16.9582	-0.03	0.2597	0.6692	0.39	9.7036	6.3365	1.53
High school graduate	-24.6122	16.6818	-1.48	2.9657***	0.6654	4.46	41.5547***	6.4365	6.46
Post-secondary/associate degree	-14.8805	18.2926	-0.81	5.1137***	0.7445	6.87	83.4660***	9.0927	9.18
College degree or higher	34.4609	21.8180	1.58	4.6625***	0.8734	5.34	163.9384***	17.9561	9.13
Race									
Native American	-17.4687	49.6692	-0.35	0.1325	1.9237	0.07	5.2426	23.7536	0.22
Asian	57.3725	56.6339	1.01	-3.8100**	1.7111	-2.23	-40.9341**	17.1552	-2.39
White	16.1171*	9.6736	1.67	-0.5949	0.4910	-1.45	-6.2492	5.1003	-1.23
Hispanic or Latino	-26.7465	26.3873	-1.01	0.2442	0.9318	0.26	-10.9634	11.0435	-0.99
Gender (Female)	8.6879	7.7448	1.12	-2.5881***	0.2936	-8.81	-32.0111***	3.7005	-8.65
Veteran	-18.0695	19.4540	-0.93	1.3458*	0.7870	1.71	29.1764**	13.6200	2.14
Primary disability									
Physical impairments	-25.4000	20.0359	-1.27	0.1200	0.8080	0.15	-2.3717	17.4014	-0.14
Mental impairments	-52.5525***	18.0453	-2.91	-2.8383***	0.7448	-3.81	-55.7796***	15.5328	-3.59
Secondary disability									

Sensory/communication impairments	-14.1650	17.5393	-0.81	-3.0087***	0.7232	-4.16	-25.6520***	8.5062	-3.02
Physical impairments	-7.8156	10.0790	-0.78	-1.2690***	0.4236	-3.00	-8.4449	5.6663	-1.49
Mental impairments	-15.1077*	8.2807	-1.82	-0.1644	0.3292	-0.50	-4.0529***	1.5932	-26.29
<b><i>Pre-service Status</i></b>									
Currently employed	7.905325	9.4227	0.84	-0.2275	0.3484	-0.65	11.1346**	5.1178	2.18
Work disincentives	3.0327	2.971615	1.02	-3.8039***	0.1238	-30.47	-41.8839***	1.5932	-26.29
Previous closure/service									
Closed before services	-13.2639	13.1365	-1.01	0.1923	0.5492	0.35	-8.0276	6.8302	-1.18
Closed after services	-22.0510***	8.3032	-2.66	-1.2896***	0.3354	-3.85	-18.4833***	3.8301	-4.83
Participation in Projects with Industry	-41.4231	59.0265	-0.70	-1.4386	2.4348	-0.59	-61.2789**	25.8489	-2.37
<b><i>Employment Services</i></b>									
No. of placement services received	93.0754***	12.3244	7.55	-1.9901***	0.4011	-4.96	-35.1571***	5.7619	-6.10
Constant	292.321***	31.31243	9.34	37.6258***	1.2273	30.66	336.2284***	19.0697	17.63
<i>F</i> -test	15.70***			81.59***			56.09***		
<i>R</i> <sup>2</sup>	.2707			.2601			.2805		

\*significant at .1; \*\*significant at .05; \*\*\*significant at .01; two-tailed tests.

## Appendix I. Covariate Balance Check Before and After Matching

The absolute standardized difference of covariate means is the absolute value of the mean difference as a percentage of the average standard deviation. For each covariate X,  $\bar{x}_t$  and  $\bar{x}_c$  are the means in the treatment and control groups, and  $s_t^2$  and  $s_c^2$  are the corresponding variances, respectively. The absolute standardized difference includes two standardized measures:

$d_x$  contrasts covariate values for treatment units with those of all control units before matching

$$d_x = \frac{|\bar{x}_t - \bar{x}_c|}{\sqrt{(s_t^2 + s_c^2)/2}}$$

$d_{xm}$  contrasts covariate values for treatment units with those of all the matched controls after matching (a subscript  $m$  for after matching)

$$d_{xm} = \frac{|\bar{x}_{tm} - \bar{x}_{cm}|}{\sqrt{(s_{tm}^2 + s_{cm}^2)/2}}$$

**(For individuals receiving employment services)**

Covariate	2004			2005			2006			2007			2008			2009		
	$d_x$ (%)	$d_{xm}$ (%)	t-statistic	$d_x$ (%)	$d_{xm}$ (%)	t-statistic	$d_x$ (%)	$d_{xm}$ (%)	t-statistic	$d_x$ (%)	$d_{xm}$ (%)	t-statistic	$d_x$ (%)	$d_{xm}$ (%)	t-statistic	$d_x$ (%)	$d_{xm}$ (%)	t-statistic
	Before matching: N <sub>IN</sub> =2951, N <sub>MI</sub> =2148 After matching: N <sub>IN</sub> =N <sub>MI</sub> =1598			Before matching: N <sub>IN</sub> =3048, N <sub>MI</sub> =1143 After matching: N <sub>IN</sub> =N <sub>MI</sub> =955			Before matching: N <sub>IN</sub> =2673, N <sub>MI</sub> =1035 After matching: N <sub>IN</sub> =N <sub>MI</sub> =852			Before matching: N <sub>IN</sub> =2770, N <sub>MI</sub> =1213 After matching: N <sub>IN</sub> =N <sub>MI</sub> =1026			Before matching: N <sub>IN</sub> =2762, N <sub>MI</sub> =1098 After matching: N <sub>IN</sub> =N <sub>MI</sub> =970			Before matching: N <sub>IN</sub> =2569, N <sub>MI</sub> =887 After matching: N <sub>IN</sub> =N <sub>MI</sub> =785		
Age	22.6	12.1**	3.39	7.0	14.4**	3.14	11.8	17.7**	3.65	15.1	5.8	1.67	22.6	7.8	1.63	30.0	11.9**	5.61
Gender	4.0	6.9	1.93	7.4	5.8	1.27	5.5	10.6**	2.17	11.3	12.2**	2.75	7.8	13.3**	2.92	2.1	0.5	0.10
Race	23.7	1.8	0.50	6.7	6.1	1.30	12.7	7.2	1.46	12.5	5.4	1.63	17.0	2.0	0.40	14.0	5.8	1.07
Education	0.1	5.0	1.39	2.8	0.6	0.13	7.1	4.5	1.22	0.5	2.9	0.63	3.4	5.3	1.11	7.4	7.6	1.50
Veteran status	22.1	15.5**	4.44	18.4	20.7**	4.25	11.5	18.4**	3.41	6.8	7.9	1.70	20.9	7.8**	3.99	14.7	8.4	2.04
Projects with industry	2.9	2.8	0.71	7.7	1.9	0.45	4.6	0.0	0.00	8.9	1.9	0.58	2.8	2.3	0.38	9.8	2.1	0.45
Primary disability	12.0	5.4	1.47	18.7	16.7**	3.59	10.7	7.3	1.53	15.1	11.4**	3.29	15	7.7**	2.15	15.3	3.4	0.88
Secondary disability	36.6	6.0	1.69	37.2	16.3*	3.53	22.9	13.3**	2.75	21.7	16.6	3.80	28.7	13.6**	2.99	25.1	9.8	1.94
Employment status	10.8	3.5	0.94	4.4	0.0	0.00	7.8	6.5	1.33	1.0	2.8	0.63	9.2	1.7	0.48	0.8	1.2	0.23
Work disincentives	1.6	0.1	0.02	5.9	1.4	0.30	11.0	4.4	0.90	0.2	8.9	1.96	4.4	4.8	1.02	0.5	2.7	0.54
Previous closure/service	11.2	3.9	1.10	8.8	9.4**	2.03	8.7	9.0	1.86	6.5	3.5	1.00	16.3	4.5	1.25	13.2	6.2	1.59
No. of placement services received	14.9	8.0**	2.40	69.9	6.5	1.49	69.9	3.0	0.67	60.9	1.8	0.44	55.5	4.1	0.93	57.2	2.7	0.56

\*\*significant at .05; two-tailed tests.

**(For individuals with employment)**

Covariate	2004 Before matching: N <sub>IN</sub> =1185, N <sub>MI</sub> =862 After matching: N <sub>IN</sub> =N <sub>MI</sub> =525			2005 Before matching: N <sub>IN</sub> =1196, N <sub>MI</sub> =611 After matching: N <sub>IN</sub> =N <sub>MI</sub> =431			2006 Before matching: N <sub>IN</sub> =1303, N <sub>MI</sub> =553 After matching: N <sub>IN</sub> =N <sub>MI</sub> =376			2007 Before matching: N <sub>IN</sub> =1429, N <sub>MI</sub> =616 After matching: N <sub>IN</sub> =N <sub>MI</sub> =445			2008 Before matching: N <sub>IN</sub> =1277, N <sub>MI</sub> =550 After matching: N <sub>IN</sub> =N <sub>MI</sub> =398			2009 Before matching: N <sub>IN</sub> =1048, N <sub>MI</sub> =404 After matching: N <sub>IN</sub> =N <sub>MI</sub> =295		
	$d_x$ (%)	$d_{xm}$ (%)	t- statis- tic															
Age	21.1	8.4	1.38	22.5	8.7	1.76	23.1	8.6	1.69	9.7	0.9	0.19	36.4	15.3**	2.97	25.0	17.8**	2.98
Gender	0.5	5.5	0.90	7.5	7.6	1.14	10.1	5.0	0.68	14.7	14.9**	2.20	11.0	3.6	0.71	4.0	5.6	0.68
Race	15.8	5.9	0.90	11.0	8.7	1.80	6.0	1.0	0.14	16.3	3.6	0.51	13.6	10.4	1.41	8.6	2.0	0.22
Education	3.7	4.7	0.73	8.9	6.0	1.39	22.2	7.0	1.34	11.7	5.1	1.05	13.2	12.7	1.70	13.4	4.6	0.54
Veteran status	25.2	12.5**	2.43	16.3	13.6**	2.96	44.1	18.2**	3.25	19.2	12.0**	2.33	4.9	3.9	0.45	7.0	9.1	1.08
Projects with industry	6.0	5.8	1.00	5.7	4.0	0.58	3.9	0.0	--	5.7	5.1	0.58	8.6	0.9	0.17	12.4	4.0	0.58
Primary disability	17.6	14.0**	2.23	23.2	18.8**	2.7	15.4	7.9	1.55	10.4	10.3	1.93	20.4	15.8**	3.03	8.2	1.1	0.19
Secondary disability	35.3	13.8**	2.21	31.1	12.9	1.91	23.9	18.3**	3.63	21.4	23.1**	3.51	25.9	7.2	1.35	33.4	15.5	1.89
Employment status	9.6	5.5	0.88	2.9	1.2	0.16	4.8	1.1	0.22	11.0	2.0	0.04	5.4	0.2	0.04	7.3	1.4	0.24
Work disincentives	0.5	1.7	0.27	12.0	7.4	1.08	17.7	13.8**	3.10	5.9	1.0	0.15	21.1	4.2	0.82	19.5	7.1	0.85
Previous closure/service	5.5	3.0	0.48	9.4	0.7	0.15	13.9	6.3	1.24	12.4	2.8	0.58	19.7	5.8	1.14	3.3	5.6	0.70
No. of placement services received	21.2	1.7	0.34	89.8	3.1	0.52	98.2	0.0	0.00	89.4	1.2	0.21	88.9	1.3	0.22	90.7	6.2	0.83

\*\*significant at .05; two-tailed tests.