7-2006

Does the Quality of Training Programs Matter? Evidence from Bidding Processes Data

Alberto Chong
Inter-American Development Bank

Jose Galdo
Syracuse University

IZA DP No. 2202

Does the Quality of Training Programs Matter? Evidence from Bidding Processes Data

Alberto Chong
Jose Galdo

July 2006

Forschungsinstitut
zur Zukunft der Arbeit
Institute for the Study
of Labor

Does the Quality of Training Programs Matter? Evidence from Bidding Processes Data

Alberto Chong
Inter-American Development Bank

Jose Galdo
Syracuse University and IZA Bonn

Discussion Paper No. 2202
July 2006

IZA

P.O. Box 7240
53072 Bonn
Germany

Phone: +49-228-3894-0
Fax: +49-228-3894-180
Email: iza@iza.org

IZA Discussion Paper No. 2202
July 2006

ABSTRACT

Does the Quality of Training Programs Matter? Evidence from Bidding Processes Data*

We estimate the effect of training quality on earnings using a Peruvian program, which targets disadvantaged youths. The identification of causal effects is possible because of two attractive features in the data. First, selection of training courses is based on public bidding processes that assign standardized scores to multiple proxies for quality. Second, the evaluation framework allows for the identification and comparison of individuals in treatment and comparison groups six, 12, and 18 months after the program. Using difference-in-differences kernel matching methods, we find that individuals attending high-quality training courses have higher average and marginal treatment impacts. External validity was assessed by using five different calls over a nine-year period.

JEL Classification: I38, H43, C13, C14

Keywords: training, quality, earnings, bidding, matching methods

Corresponding author:

Jose Galdo
Center for Policy Research
Syracuse University
Eggers Hall
Syracuse, NY 13244-1020
USA
E-mail: jcgaldo@maxwell.syr.edu

* Jeffrey Kubik and Hugo Ñopo provided valuable comments. Arturo Garcia provided excellent research assistance. The views and opinions in this paper should not be attributed to the Inter-American Development Bank or its Executive Directors. The standard disclaimer applies.

1. Introduction

Despite the fact that the empirical evidence on active labor policies suggest that training programs for the youth and displaced are not worth the cost, such programs keep being reinvented by policymakers. This has been particularly true in recent years where massive privatization processes and dramatic reduction in overstaffed public sectors have driven a large fraction of workers to the unemployment ranks or the underground economy. In fact, training programs appear to yield small and even negative returns in both developed and developing countries (Heckman, Lalonde, and Smith, 1999; Rama, 1999).[1] In this context, it is by no means clear whether training programs are ineffective because they target relatively unskilled and less able individuals or simply because of quality issues. After all, the same government agencies that get low grades in training assessments are the ones that end up in charge of the training component of downsizing operations.[2]

The evaluation of training programs has played a central role in studying the effectiveness of active labor-market policies. The paradigm of the representative agent that assumes that a public policy has the same impact on all treated individuals has produced a vast array of empirical work that focuses on mean treatment impacts. Theory, however, predicts systematic heterogeneity in the impact of active labor-market programs on earnings and income (e.g., Bitler, Gelbach, and Hoynes, 2004). In particular, whereas a growing literature focuses on the causal effect of variations in school or college quality on earnings (e.g., Black and Smith, 2003, Dale and Krueger, 2002, Card and Krueger, 1992), evidence for training programs is non-existent.

In this paper, we study the link between earnings and quality of training services. To our knowledge, this is the first paper that addresses quality issues in training programs, with the added advantage that we are able to use disaggregated data at the course level, rather than at the school or state level. In fact, the selection of training

---

[1] For instance, Campa (1997) shows the limited ability of training programs to reallocate workers to alternative industries, partly because training was focused on the update of previous skills rather than the acquisition of new ones.

[2] An implicit assumption is that productivity-enhancing effects of quality training explain higher labor earnings. This may be due to human capital accumulation as better-trained individuals are more productive and, as a result, obtain higher earnings. Alternatively, since the cost of acquiring training is lower for high-ability individuals, even if training is unproductive, firms may make inferences about productive differences from training choices and workers respond by selecting longer training to signal higher quality. For our purposes, both models yield similar empirical predictions.

courses is based on public bidding processes where a small team of education specialists assigns standardized scores to multiple proxies for quality. This allows us to provide meticulous focus on within-school variation rather than on between-schools, which may improve the explanation for quality heterogeneity (Hanushek, Kain, O'Brien, and Rivkin, 2005). Furthermore, the detailed bidding questionnaires and instruments not only targets common proxies for quality such as expenditures per student, class size, infrastructure, equipment, and teacher schooling, but also emphasized teacher skills and curricular structure, such as the consistency of goals, contents, and activities, which may improve the predictive power of the quality measurement.

This paper takes advantage of a non-experimental program, the Peruvian Youth Training PROJOVEN Program, which has provided training to around 35,000 disadvantaged young individuals aged 16 to 25 since 1996. The program design has changed the government's intervention in the training markets from unconditional funding to public institutions towards conditional cash transfers to public and private institutions competing for restricted public funds. The treatment consists of two sequential and articulated phases at the training institution (formal classes) and at productive firms (on-the-job-training) for a period of six months. To guarantee the paid, on-the-job training experience for each trainee, the program's design follows a demand-driven approach where competing institutions have to structure the training courses in accordance with the labor demand requirements of productive firms.[3] Hence, this unique data design allows us to examine the effectiveness of market-based approaches in the provision of training services. Furthermore, the evaluation framework allows us to identify and compare treatment and comparison group individuals six, 12, and 18 months after the program, which also allow us to test the sustainability of treatment impacts across time.

The comparison group individuals are selected from a random sample of "nearest-neighbor" households located in the same neighborhoods of those participants included in the evaluation sample. This costly evaluation design greatly ameliorates support problems in the data, which is one of the most important criteria needed for solving the evaluation problem. Indeed, both the standardized quality scores based on bidding

[3] Similar programs were implemented since mid-1990s in Chile (Chile Joven), Argentina (Proyecto Joven), and Colombia (Youth Training Program).

information and the unique evaluation framework allow us to overcome two crucial problems frequently encountered in the literature, data limitations and econometric problems, and to provide alternative measures to typical point estimates which have been highly criticized (Glewwe 2002).

To the extent that socioeconomic variables and family background raise or lower earnings for all levels of training attainment, we merge the bidding data to the baseline and follow up evaluation data that contain individual and household information for both treatment and comparison groups. Thus, our estimates treatment impacts are purged of any effects coming from demographic, socioeconomic, and family differences. Furthermore, the availability of data for five different cohorts of individuals over a nine-year period (1996 to 2004) allows us to consider the robustness of our estimates with respect to the external validity assumption.

A related problem that is difficult to overcome in conventional studies is unobserved characteristics of individuals and households that cause omitted variable bias in OLS estimations. With few exceptions, this literature relies on what Heckman and Robb (1985) call “selection on observables” to identify the effects of education quality on labor outcomes. To control for potential bias arising from differences in observed and unobserved characteristics, difference-in-differences kernel regression matching methods are implemented that allow for selection on time-invariant unobservables (Smith and Todd, 2005). We also implement an alternative marginal matching estimator that assumes that sorting into different quality training courses arises from both observables and unobservables. An advantage of this estimator is that it only requires data for the treatment group and thus can be implemented when no comparison group data are available (Behrman, Chen, and Todd, 2004).

Our empirical findings can be summarized in four conclusions. First, we find strong evidence about the effectiveness of market-based approaches in the provision of training services. In fact, the combination of bidding processes with demand-driven approaches that ensures quality and pertinence of the training courses yield larger overall point estimates than those reported in the literature. This result is particularly robust for females who show much higher treatment impacts than male participants. Second, we find evidence of substantial heterogeneity in response to training quality. In general, individuals attending high-quality training courses show labor earnings 20 percentage points higher than individuals attending low-quality courses. Third, the marginal returns to training quality are higher for those individuals who complete only the first stage of the program at the training centers rather than that for those individuals who complete both stages of the program. This result indicates that the second stage of the program, the on-the-job-training experience, smooth productive gains between people attending high- and low-quality training courses. Fourth, the returns to training quality are not steady across time: the earnings gap between individuals attending high- and low-quality courses is higher in the medium-term than in the short-run, even though the average returns to training depreciate over time.

The remainder of the paper is organized as follows. In section 1 we provide an overview of the PROJOVEN program. We then discuss the measurement of training quality in section 2. In section 3 we present the evaluation data. In section 4 we discuss the empirical strategy along with the identification assumptions. Our main results appear in section 5. In section 6 we show some robustness tests, and we conclude in section 7.

2. The PROJOVEN Program

To smooth the short-run negative effects of structural reforms on the welfare of poor households in Latin American during the mid-1990s, several countries launched active labor-market policies. In particular, the disproportionately large unemployment rates for young individuals galvanized the implementation of training programs across the region. The most distinctive element differentiating this generation of training programs from previous experiences was the use of market-based approaches intended to promote competition among public and private training institutions.

The Youth Training Program, PROJOVEN, was implemented in 1995 with the goal of increasing the employability and productivity of disadvantaged young individuals aged 16 to 25 via basic training courses. The treatment consists of a mix of formal and practical training divided into two phases. The first stage consists of 300 hours of formal classes at the training center locations -roughly five hours per day for three months. In the second phase, training institutions must place trainees into a paid, on-the-job training experience in productive firms for an additional period of three months.

The success of this program design presumes a strong match between the content of the courses in the first stage and the firm’s labor skill requirements. The program operator enhances this match via a demand-driven design where the competing training institutions must show formal agreements signed with private productive firms guaranteeing three-month on-the-job training experience for the beneficiaries. In addition, the quality of training services is highly considered by the program operator, who implements public bidding processes in order to choose the highest-quality training services at the best competing prices. Conditional payments based on the training centers' effectiveness in ensuring completion of the six-month course provide incentives to these institutions to implement high-quality courses.[4]

2.1 The Beneficiary Selection Process

PROJOVEN's selection process consists of multiple stages governed by different actors: target individuals, bureaucrats, and training centers. Figure 1 shows the dynamic of this process. The program awareness strategy (position A) constitutes the first formal effort to reach out to the target population and aims to inform potential participants about the program's benefits and rules. This first filter focuses only on those neighborhoods with a high concentration of households below the poverty line. Those prospective participants attracted by the expected benefits and perceived opportunity costs of participation voluntarily show up in the registration centers (position B) where qualified personnel determine their eligibility status. A standardized targeting system based on five key observable variables (poverty status, age, schooling, labor market status, and pre-treatment earnings) determines who is eligible and who is not. This process concludes when the total number of eligible individuals exceeds by around 90 percent the total number of slots available in each call.

A two-tiered monitoring and supervision process guarantee the reliability of the information given by the prospective applicants to determine their eligibility status. In addition to focusing only on poor targeted districts, the program operator makes house visits to those applicants who provided dubious or inconsistent information. Finally, a random sample of eligible and non-eligible individuals is subject to an ex-post visit, which allows the program operator to detect misreported cases and improve the eligibility survey and instruments.

The eligibility status does not guarantee participation in the program. Program enrollment depends on both training centers' and applicants' willingness to pursue the

[4] Payments are structured in per capita terms according to the following scheme: 100, 80, 60, and 30 percent if completing six, four, three, and one month of training, respectively.

application process to its conclusion. Eligible individuals are invited to an orientation process (position C), where they choose the courses they want to attend following a first-come-first-serve criterion. This process concludes when the number of eligible individuals exceeds around 75 percent of the number of available slots for each course.

Finally, the training institutions select beneficiaries from the pool of eligible applicants generated by the program (position D). This final step does not follow standardized criteria since each institution applies its own rules. It is important to note that because the eligible/beneficiary ratio is around 1.75, the role of the training centers in selecting the beneficiaries is limited.[5]

3. Measuring Training Quality

The selection of training services follows a two-step standardized process. The first step targets the selection of training institutions. The program operator consults a training directory called RECAP, which list all the training institutions eligible to participate in the program. To be included in the RECAP, the training centers must pass a minimum quality threshold following standardized instruments that mostly evaluate their legal status (formality) and the existence of some acceptable level of human resources and infrastructure. In this step, institutions do not compete with each other, and there are no restrictions as to the number of institutions that can be listed in the RECAP.[6]

In the second step, the program operator invites institutions included in the RECAP to participate in public bidding processes where the selection of training courses rather than training institutions takes place. A formal and blind evaluation guarantees the transparence of this process where a small team of education specialists carefully evaluates the proposals of each competing course according to a battery of standardized instruments. Because of tight government budgets, the program operator selects those courses with the relative highest scores at the best competing prices.[7]

Three distinctive features characterize the quality measurement in this second step. First, all proxies for quality are disaggregated at the course level rather than at

[5] It is against the program's rules to select individuals based on age, race, sex, and schooling. From interviews with both the program operator and training institution personnel, it seems that the final selection of beneficiaries is driven by variables such as marital status, children, and specific physical requirements arising from the courses (e.g., body mass for handling weights).

[6] For those training centers that participate in two or more consecutive calls, the previous performance is also considered as an additional evaluation factor. It explains almost half of the total score.

[7] The number of selected courses depends on the available training slots that are determined ex-ante.

school level, which allows us to measure the quality of the training services in great detail. Thus, variations can be found within training centers depending on the relative distribution of school supplies or differential teacher experience across courses.

Second, detailed questionnaires and instruments not only target common proxies for quality such as expenditures per student, class size, infrastructure, and equipment, but also put emphasis on the curricular structure (i.e., consistency among goals, contents, and activities) and teacher "skills" (i.e., experience in dealing with disadvantaged young individuals). The inclusion of this new set of "soft" variables that defy an objective description may dramatically improve the explanation for differences in quality (Hanushek, 1986).

Third, the measurement of quality proxies follows a standardized system of scores rather than the classical approach of computing raw quantities (e.g., number of computers). In this way, the evaluators are able to evaluate both the number of items in each subcategory and their intrinsic quality. For example, in evaluating a course on computing software, the total score in the equipment variable will depend on both the quantity of computers per student and the model and antiquity of the machines. The use of standardized scores also allows for the evaluation of variables such as curricular structure that do not per se have a corresponding quantitative content. Only two proxies for quality are measured in raw form: expenditures per student and class size.

This paper focuses on six different categories of proxies for quality: class size, expenditures per trainee, eight teacher variables, six infrastructure and equipment physical characteristics, nineteen curricular structure variables, and nine variables characterizing the link between the content of the training courses and the institution's knowledge about workers and occupational analysis of labor demand. As a whole, these variables largely exceed the number of school and teacher characteristics considered to b core variables in the literature (Fuller 1987; Harbison and Hanushek, 1992). Table A.1 in Appendix describes the full set of variables.

Table 1 displays summary statistics of these quality measures using re-scaled indices for all categories. We use data from 1996 to 2004, which allows us to identify five different bidding processes corresponding to the first, second, fourth, sixth, and eighth calls of the program. Two features emerge. First, there is variation in the scores assigned to each category within and across calls. In particular, expenditures per trainee and curricular structure are the variables that vary most across institutions and calls. On the other hand, infrastructure and equipment are the variables that show the smallest variation. Second, as one might expect, there is an increasing trend in the average quality for some proxies over time. This is explained by a natural learning curve on the part of continuously participating institutions, and by the program operator's inability to attract a large number of new training institutions.[8]

The lower panel of Table 1 combines the information for all quality proxies using factor analytic methods to produce a one-dimensional "quality index". In doing so, we use the first principal component that is a linear combination of the quality proxies that accounts for the highest proportion of their variance. Subsequent principal components are orthogonal linear combinations that explain the highest proportion of the remaining variation. Thus, we report only results based on first factors in all cases. We observe large variability in the index within and across calls, which may play an important role in explaining heterogeneous treatment impacts in the program.

We also include the number of competing training institutions, courses offered, and courses accepted for these five calls. The average number of training institutions is 33 per call, ranging from 30 to 48. These institutions offered an average of 200 courses per call. We also observe that the supply of training courses and the number of selected curses have followed parallel paths. The ratio of funded courses / competing courses reaches 0.59, which indicates a relatively high probability of success for those training institutions included in the RECAP.

Two potential factors that may affect the accuracy with which the quality proxies are measured are evaluation bias and misreporting. In the first case, evaluators may introduce bias when assigning scores due to subjective evaluation. The program operator, however, minimizes this risk by hiring and training a small team of education specialists who are trained to follow a standardized score system. The competition for limited public resources may also encourage training centers to misreport public offers. To minimize this problem, the program operator has implemented a monitoring system that uses inspections to ensure the validity of all technical specifications contained in the offers.

[8] The average number of new training institutions entering in successive calls is 9. Indeed, this number reveals that failure of one of the long-term objectives of the program: the generation of a more competitive training market.

The bidding data are then merged to the evaluation data, which imply that all treated individuals attending the same training course receive the same quality scores.

4. The Evaluation Data

The PROJOVEN evaluation datasets consist of panel data collected in four rounds on 10 different sub-samples associated with five different calls (cohorts) of beneficiaries receiving treatment between 1996 and 2004, and five corresponding comparison group sub-samples. The panel data consist of a baseline and three follow-up surveys taken six, 12, and 18 months after the program. The beneficiary sub-samples are selected from a stratified random sample of the population of participants corresponding to the first, second, fourth, sixth, and eighth call of the program.

Individuals in the corresponding comparison sub-samples are selected from a random sample of "nearest-neighbor" households located in the same neighborhood as those participants included in the evaluation sample. The program operator builds the comparison samples by using the same eligibility instruments applied to the treatment sample and pairing each beneficiary to a random neighbor that has the same sex, age, schooling, labor market status, and poverty status. The neighborhood dimension has the ability to control some unobservables, including geographic segregation, transportation costs, and firms' location, which may affect propensity to work and the potential outcomes.

The baseline databases provide rich information about individual socioeconomic and labor-market characteristics, parental characteristics, and dwelling characteristics. In fact, relevant factors affecting both the propensity to participate in the program and labor market outcomes are available. There is information, for example, on education attainment, marital status, number of children, parents' schooling, and participation in welfare programs. The labor-market module includes information about working status, experience, monthly earnings, type of work, firm's size, and participation in previous training courses. At the household level, the data sets provide information about family size, family income, and household's density rate. In addition, the datasets provide detailed information on dwelling characteristics including source of drinking water, toilet facilities, and house infrastructure (type of materials used in the floor, ceiling, and walls).

4.1 Comparison of Pre-Treatment Sample Means

Table 3 compares the means of several covariates for the treatment and comparison samples for each one of five different cohorts. Column 2 shows the means using the pooled sample and columns 3 to 7 show the p-values for test of difference of means for each cohort. In terms of demographic and socioeconomic characteristics, Panel A shows the effectiveness of the "neighborhood" strategy to balance the distribution of covariates used to determine eligibility status. Both groups have the same average age (19), sex ratio (42 percent are males), and schooling attainment (85 percent have completed high school). The p-values for all cohorts under analysis do not reject the null hypothesis of equality of means. The data shows, however, that both marital status and children variables have different distributions. About 90 percent of the participants are single and only 14 percent have children, which differ from the comparison sample, which has a lower proportion of single people (77 percent) and higher proportion of individuals with offspring (25 percent). The p-values show that this is a robust result for all cohorts.

Panel B compares labor market characteristics for treatment and comparison samples. Both groups have the same proportion of individuals in and out of the labor force. Approximately 52, 25, and 22 percent of individuals were employed, unemployed, and out of the labor force, respectively. These non-significant differences are consistent across all cohorts as is shown by the p-values. The type of work depicts a somewhat different pattern. A higher proportion of comparison individuals were working in the formal private sector (63 versus 54 percent) whereas a higher proportion of treated individuals were non-paid family workers (17 versus 10 percent). A comparison of monthly earnings also shows that treated units receive on average smaller earnings than their counterpart comparison sample, which is a steady result across all cohorts.

Panel C compares households and dwelling characteristics. On average family income are somewhat smaller for treated individuals, although the p-values show mixed results across different cohorts. In addition, the analysis of dwelling characteristics shows that a higher proportion of treated individuals live in houses with somewhat better infrastructure and access to flush toilet and piped water. These differences, however, are not significant for several cohorts. Finally, Panel D shows parental schooling attainment. In general, the schooling distribution in both samples is similar, with mothers having fewer years of formal education than their spouses. The p-values do not reject the null hypothesis of equality of means for most of the categories.

In summary, the baseline datasets show that we are dealing with a homogenous population in terms of several socioeconomic and labor-market characteristics, including sex, age, schooling, parents' education, type of work, previous training, and family size. On the other hand, the data also reveal some significant differences in variables such as marital status, children, monthly earnings, family income, and some dwelling characteristics, which would play an important role in our any econometric strategy intended to eliminate selection bias.

5. The Empirical Strategy

Let $Y_1(q)$ be the potential outcome in the treatment state $(T=1)$ for an individual who participated in a training course of quality $q=\{q_1,q_2,...\}$ and let $Y_0(q')$ be the potential outcome in the untreated state $(T=0)$. In our application, the untreated state refers to either no participation in the program, in which case $q'=0$, or participation in a training course of quality q', where $q'<q$. We observe the pairs $(Y_1(q),T_1)$ and $(Y_0(q'),T_0)$ but never $(Y_1(q),T_0)$ or $(Y_0(q'),T_1)$. Because of this missing data problem, we cannot identify for any particular individual the treatment gains $\Delta_i=(Y_1(q)-Y_0(q'))$. We focus, instead, on both average and marginal treatment impacts conditional on the quality of the training courses.

Our parameter of interest is the impact of treatment on the treated that estimates the mean effect of attending a high-quality training course rather than not participating (or attending a low-quality course) on the individuals who attend a high-quality course:

$$\Delta_{TT}=E(Y_1(q)-Y_0(q')\,|\,T=1)=E(Y_1(q)\,|\,T=1)-E(Y_0(q')\,|\,T=1). \qquad (1)$$

While $E(Y_1(q)\,|\,T=1)$ may be estimated from the observed treatment sample, the right-hand side of the equation (1) contains the missing data $E(Y_0(q')\,|\,T=1)$. Using the outcomes of untreated individuals to approximate the missing counterfactual yield the well-known selection bias because of differences in the distribution of observed and unobserved characteristics between *T=1* and *T=0.*

To eliminate the selection bias, we implement matching methods to estimate the counterfactual outcome for program participants by taking weighted averages over the outcomes of observationally similar untreated individuals. Thus, we relax any linear assumption that may mask the earnings-quality relationship.[9] We proceed under the assumption that the distribution of unobservables varies across *T=1* and *T=0* but not over time within groups, which is the standard assumption of difference-in-differences models.

5.1 Identifying program impacts when the counterfactual is not participation ($q'=0$)

In general, standard matching methods eliminate selection bias by balancing the distribution of observables of the untreated group with that of the treated group. However, there may be systematic differences in *T=1* and *T=0* outcomes even after conditioning on a rich set of observables. Such differences may arise in the PROJOVEN program from three different sources. First, it is impossible to control differences in innate ability or motivation.[10] Second, we do not observe all the factors that govern the transition from eligible status to beneficiary status. Third, we may not observe and measure certain aspects of teacher and school quality correlated with the quality index.

To eliminate bias arising from unobservables, we can use difference-in-differences matching methods (Heckman, Ichimura, and Todd, 1997) that are conditional semiparametric versions of the widely used parametric approach. This method solves the evaluation problem by subtracting the before-after change in untreated outcomes from the before-after change for treatment outcomes. The identifying assumption justifying this matching estimator is that there exists a set of conditioning variables *X* such that

$$E(Y_t(q)-Y_{t'}(q')\mid X,T=1)=E(Y_t(q')-Y_{t'}(q')\mid X,T=0)\,. \qquad (2)$$

where *t'* and *t* refer to before and after the start of the program and $q'=0$. This assumption ensures that after conditioning on a rich set of observable variables, the outcomes for treated and untreated individuals follow a parallel path. Put differently, conditional on *X*, participation in a training course of quality *q* is unrelated to what your mean growth outcome would be if one did not participate in the program.

Matching methods force us to compare comparable individuals by relying on the common support assumption

$$\Pr(T=1\mid X)<1 \text{ for all } X. \qquad (3)$$

[9] For instance, Heckman, Layne-Ferrar, and Todd (1995) find that estimated earnings-quality relationships are sensitive to specification of the earnings function. When false linearity assumptions are relaxed, the only effect of measured schooling quality is on the returns for college graduates.
[10] We do not have information about any IQ or Raven's matrices tests.

The support condition ensures that for each X satisfying assumption (2) there is a positive probability of finding a match for each treatment individual. Otherwise, if there are X for which everyone received treatment, then it is not possible for matching to construct the counterfactual outcomes for these individuals. In this sense, matching forces us to compare comparable individuals in a way that standard regression methods do not.

Under conditions (2) and (3), we estimate the treatment impacts by computing first the counterfactual outcome for each treatment unit using a weighted average of the comparison units' outcomes over the common support region, and then averaging these results over the treatment group sample

$$\Delta^{DID} = \frac{1}{n_1} \sum_{i \in n_1 \cap S_p} \left\{ [Y_t(q) - Y_{t'}(0)] - \left\{ \sum_{j \in n_0 \cap S_p} W(\| i - j \|)[Y_t(0) - Y_{t'}(0)] \right\} \right\}. \quad (4)$$

where n_1 and n_0 are the sample of treatment and comparison individuals, S_p is an indicator function that takes the value 1 for individuals in the common support region (0 otherwise) and $W(\| i - j \|)$ is the key weighting function that depends on the Euclidian distance between each comparison group individual and the treatment group individual for which the counterfactual is being constructed. We estimate the counterfactual outcome $\sum W(\| i, j \|)[Y_t(0) - Y_{t'}(0)]$ using local linear regression methods that were developed in the early 1990s by Fan (1992) and have more recently been considered in the evaluation literature by Heckman, Ichimura, Smith, and Todd (1998). This nonparametric approach relies on standard kernel weighting functions that assign greater weight to individuals that are similar, and is more efficient than local constant regression methods because of its lower boundary bias in regions of sparse data.

5.2 Identifying marginal program impacts $(q' < q)$

We are also interested in the marginal treatment impacts of increasing quality in the program from q' to q, where q'>0, using data on program participants who have received different qualities of treatment. An important advantage of using only treatment individuals is that we do not require assumptions about the process governing selection into the program. On the other hand, this approach may introduce a potential source of nonrandom selection because of potential sorting. Indeed, this is the main econometric problem in studies addressing the link between college quality and labor earnings.[11]

The rules and procedures governing the selection into the PROJOVEN program, however, severely limit the possibility that high-ability individuals select into high-quality courses. First, a relatively small sample of individuals similar in multiple dimensions is determined by standardized instruments rather than heterogeneous large populations extracted from national education surveys. Second, and most importantly, the eligible individuals choose the course they want to attend according to a first-come-first-serve criterion, which is not the case in studies addressing college education. Third, there is no evidence that training institutions use any sort of IQ tests to select the program's beneficiaries among the eligible population. Fourth, even if the training institutions select the smartest eligible individuals, high- and low-quality institutions have the same chances of sorting because the program operator sends them the same number of eligible individuals. Finally, the ability of the training institutions to select a "smart" group is very restricted because only a relatively small proportion of eligible individuals is rejected.

Because we cannot discard some sort of sorting in our data, as is shown in Columns 9 and 10 of Table 2, we again implement difference-in-differences matching methods that assume selection in observables and unobservables to eliminate selection bias. Formally, the identifying condition (2) changes to

$$E(Y_t(q_1) - Y_{t'}(0) \mid X, T=1) = E(Y_t(q_2) - Y_{t'}(0) \mid X, T=1). \qquad (5)$$

which states that the mean outcomes for individuals participating in high-quality courses follows a parallel path with respect to individuals attending low-quality courses. We estimate the marginal treatment impact using the same matching estimator (equation 4), although this is adjusted for the changes implied in assumption (5). This new estimator gives the impact of increasing the quality of the program from q_1 to q_2 for the group of individuals who enrolled in the training course of quality q_1, where $q_1 > q_2$.

[11] Recent work includes Black and Smith (2003), Dale and Krueger (2002), Brewer, Eide, and Ehrenberg (1999).

5.3 Empirical Issues

Because the "curse of dimensionality" arises when *X* is high dimensional, we follow the celebrated result of Rosenbaum and Rubin (1983), who show that if the information set contained on *X* justify matching, then matching on the balancing score $b(X)$ is also justified. The balancing scores is a function of attributes at least as "fine" as the valued index that predicts the probability of participation: the propensity score.

The proof that assumptions (2) and (5) hold for $P(X)$ instead of *X*, is attained by a balancing property,

$$E(X \mid T=1, P(X)) = E(X \mid T=0, P(X)) = E(X \mid P(X))$$

This is a non-trivial property because in general $P(X_m) \approx P(X_n)$ does not imply $X_m \approx X_n$, and hence, $E(X_m) \neq E(X_n)$.[12] Intuitively, it is equivalent to saying that conditional on $P(X)$, any differences between subgroups with different *X* are balanced out when constructing the estimates because subgroups with values of *X* that imply the same probability of participation can be combined as they will always appear in the treatment and (matched) comparison groups in the same proportion. As a result, we assume that equations (2) and (5) hold when we replace X by $P(X) = \Pr(T=1 \mid X)$.

In the empirical work, we estimate the propensity score using a maximum likelihood method (probit) and implement the balancing test suggested by Dehejia and Wahba (1999).[13] Table 3 shows the probit results for all cohorts. As expected, the covariates used to construct the comparison samples (age, sex, schooling, and work status) are not significant predictors for program selection as they are balanced between treatment and comparison groups. In general, past earnings, experience, type of work, dwelling characteristics, mother's education, family income, and family density rate are the most important predictors of participation in the PROJOVEN program. The estimates also show that married individuals and people with offspring are less likely to participate, although the coefficients are not significant for some cohorts. Furthermore, the

[12] This is a key difference with covariate matching where $X_m = X_n$ automatically implies $E(X_m) = E(X_n)$ for treatment and comparison samples.

[13] This test considers valid any parametric models that balance the distribution of pre-treatment covariates between matched individuals conditional on the propensity score. It is important to indicate, however, that multiple versions of the balancing test exist in the literature, and little is know about their statistical properties or the relative efficiency among them.

distributions of the estimated propensity scores indicate no support problems in our data. Less than 5 percent of the observations are out of the empirical overlapping region, which illustrates the relative efficiency of constructing comparison groups among eligible "neighbor" individuals. In this respect, our data satisfy one of the most important criteria needed for solving the evaluation problem.[14]

To implement the local linear kernel matching (equation 5) we also need to compute kernel functions along with their optimal bandwidths. We adopt the unbounded Epanechnikov kernel and choose bandwidth values by weighted least squares cross-validation method (Galdo, Black, and Smith, 2006), which selects the value that minimizes the mean square error of the local linear regression estimator over a bandwidth search grid. The weights account for the location of the treated units because precise estimation of counterfactuals in regions containing much of the probability mass of the treatment group individuals is more important than in regions where few treated individuals are located.[15]

6. Matching Estimates

Before presenting average and marginal matching impacts, we first report simple local linear regression estimates for the pooled sample of treatment individuals in Figure 2. The dependent variable is (real) monthly earnings that are regressed on the quality index variable using Epanechnikov kernel functions. By looking at the estimates, we observe that the quality of the training services matters. Trainees attending high-quality courses have higher labor earnings than those attending low-quality courses. Figure 2 also shows that medium-term treatment impacts (12 and 18 months after the program) are stronger than the corresponding short-run impacts (six months after the program). Because this first approach does not control any potential source of non-random selection into the program, these estimates prove only a positive correlation rather than a causal relationship.

Table 4 presents matching estimates applied separately to each one of five different cohorts. Each column refers to each cohort, and the last column shows the

[14] We follow the "trimming" method (Heckman et. al., 1998), which seems to be more stringent than alternative approaches suggested in the literature. Hence, we estimate the propensity score density distributions for $T=1$ and $T=0$ using Epanechnikov kernel functions. Then, the estimated densities are evaluated at all observed data points and, all points with zero density and points corresponding to the lowest 2 percent of estimated density values are trimmed.

[15] The bandwidth grid is defined over values 0.8 through 8 for the logs odd ratio, with a step size of 0.1.

pooled data estimates that are weighted average estimates from the five cohort samples. The upper panel (A) depicts short-run treatment impacts whereas the lower panels (B and C) present medium-term impacts. Within each panel, three different parameters of interest are presented: the average treatment effect on the treated, the average treatment effect on those attending a high-quality course, and the average treatment effect on those attending a low-quality course. In all three cases, we estimate the counterfactuals using the comparison group sample. The point estimates for the treatment impacts are presented along with their corresponding bootstrap standard errors (in parenthesis) and percentage gains (in brackets), which are calculated using the population mean earnings in the baseline period.[16]

By looking at the first row of each panel, one can observe that the PROJOVEN program is a successful, active labor market initiative as was previously shown in partial evaluations of the program.[17] The treatment impacts on the treated are positive for all cohorts, ranging in size from 33 to 70 percent six months after the program, from -5 to 89 percent 12 months after the program, and from 20 to 61 percent 18 months after the program. As expected, we also observe that the short-time treatment impacts are larger than estimates emerging 12 and 18 months after the program. For instance, by looking at the pooled sample, we see that the overall treatment impacts vary from 52 percent six months after the program to 38 and 42 percent 12 and 18 months after the program, which suggests a natural “depreciation” rate for the acquired skills.

The second and third rows within each panel show the average treatment impacts for those attending high- and low-quality training courses. In general, the matching estimates confirm the visual assessment of Figure 2: trainees attending high-quality courses have higher labor-market earnings than those trainees attending low-quality courses after purging any systematic differences in observed and unobserved covariates affecting the potential outcomes. By looking at the pooled sample estimates, we observe that the average treatment impacts for those attending high-quality courses are 58, 46, and 63 percent after six, 12, and 18 months of participation in the program. On the other

[16] Relative to their frequency in a random population, the treatment group individuals are oversampled. Because the estimated (logs) odds ratio $\hat{P}(X)/1-\hat{P}(X)$ is a scalar multiple of the true (logs) odds ratio, we apply matching methods to choice-based sampled data by using the matching variable $\hat{P}(X)/1-\hat{P}(X)$ instead of $\hat{P}(X)$.

[17] Galdo (1998), Ñopo, Saavedra, and Robles (2001), Chacaltana and Sulmont (2002).

hand, the treatment impacts for those attending low-quality courses reach 43, 24, and 24 percent, respectively. The fact that trainees attending low-quality courses also have positive treatment impacts may suggest the success of market-based demand-driven approaches that ensure the provision of selected training services.

A second important result is that the quality premium increases over time even though the average return to training depreciates across time, although most of the medium-term estimates are not statistically significant. We observe for the pooled data that six months after the program the differential effect between high and low-quality courses reaches 15 percentage points, which increases to 22 and 39 percentage points 12 and 18 months after the program.

Table 5 presents the marginal matching estimates in parallel format to Table 4.[18] Thus, we show short-run (upper panel) and medium-term (lower panels) treatment impacts for each cohort (columns 2 to 5) and the pooled data (column 6). Within each panel, we present two marginal treatment impacts: the effect of increasing the quality of the training services from q_1 (lowest quartile) to q_4 (top quartile), and the effect of increasing quality from q_3 (quartile 3) to q_4 (top quartile). Three main patterns emerge.

First, the marginal impacts indicate mostly positive treatment impacts for those attending high-quality courses. As expected, the estimates are smaller than the average treatment gains that emerge when we use non-participants as the counterfactual group and they lose statistical significance due to sample size issues. Second, the marginal impacts confirm our previous finding about the increasing quality premium over time. Thus, the estimates from the pooled sample reveal that the effect of increasing quality from q_1 to q_4 changes from 20 to 39 percent when one moves from six to 18 months after the program, and from 9 to 38 percent when one moves from q_3 to q_4. Third, by looking at the pooled data we observe that marginal treatment impacts are monotonically increasing along the quality dimension. For instance, six months after the program, the marginal impacts for those moving from q_1 to q_4 is 20 percent for the pooled sample whereas the impacts for those moving from q_3 to q_4 reaches only 9 percent.

[18] We match on the predicted probability of attending s high-quality training course (top quartile). These propensity score models are not reported but they are available upon request.

When the estimates from Tables 4 and 5 are taken together, three important lessons emerge. First, market-based approaches that put great emphasis in the quality and pertinence of the training courses yield larger overall point estimates than those reported in the literature. These large average gains are heightened by the fact that this program targets relatively skilled individuals and the per-capita expenditures on participants are not small relative to the deficits that these programs are being asked to address.[19] Thus, the PROJOVEN program has the ability, for instance, to reallocate workers from unproductive jobs (e.g., unpaid family work) toward productive ones in firms protected by international laws that guarantee minimum work conditions.[20] Second, simple average treatment impacts hide important distributional gains due to heterogeneity in the quality of the training services even within a selected group of courses that pass some quality criteria. This result suggests that the earnings gap between high- and low-quality courses would be higher if the program operator does not consider the quality dimension in the selection of the training services. Finally, this study shows the importance of having multiple cohorts when analyzing the effectiveness of active labor market programs. The sensitivity of some estimates to the sample used is significant.

7. Robustness Checks

7.1 Quality "dose" versus Treatment "dose"

The estimates for the returns to training quality may also be interpreted as returns to treatment "dose" rather than quality "dose", because of differences among training institutions to place trainees on the second stage of the program (on-the-job training experience). If this is true, it may hamper the causal relationship we have been testing in this paper. To address this potentially confounding factor, we use two different approaches. First, we check whether individuals enrolled in high-quality courses have larger treatment "doses" than individuals enrolled in low-quality courses. Using the

[19] 9 out of 10 beneficiaries have concluded high school. Also, whereas the Peruvian public school system spends S./ 470 soles per-capita each year, the PROJOVEN's per-capita expenditures reach around S./ 2,400 soles.

[20] We illustrate this fact by estimating conditional probabilistic models that use firms' size as the dependent variable (1 if working in firms with more than 20 workers, 0 otherwise) and treatment status as the key independent variable. The estimates show that treatment group individuals are 70, 52, and 47 percent more likely to work in medium- and large-size firms than comparison group individuals six, 12, and 18 months after the program. These estimates for the pooled data are statistically significant at 1 percent. It is important to note that the distribution of treated individuals across firm size is symmetric for individuals in the top and lowest quartiles of quality index.

pooled sample, we find a slight difference in favor of individuals attending low-quality courses. Over 98 percent of trainees enrolled in both low- and high-quality courses complete at least the first stage of the program, whereas 67 and 63 percent complete at least a month of the three-month on-the-job training experience.

Table 6 presents a second, more stringent test. We estimate both average and marginal treatment impacts to the subset of individuals that complete the training course at the training center location but do not participate in the paid, on-the-job practical experience. In this way, we hold fix the treatment "dose" and, at the same time, we eliminate any potential effects arising from differences among productive firms that may mask the causal effect of the training quality. We use the same matching methods and definitions as in Tables 4 and 5. The upper and lower panels show average and marginal treatment impacts for the pooled sample. Each row describes a different parameter of interest and each column refers to impacts six, 12, and 18 months after the program.

Three patterns emerge. First, the average treatment effects on the treated are positive although smaller with respect to the overall program impacts. They range from 12 percent (12 months after the program) to 28 percent (18 months after the program). Second, when considering the quality factor, the mean treatment impacts increase significantly for those attending high-quality courses. For instance, half a year after the program, the mean treatment impacts for individuals attending high-quality courses is 55 percent, which reduces to 3 percent for individuals attending low-quality courses. Third, by looking at the marginal treatment impacts, we reaffirm the strong impact of training quality on labor earnings. For the subsample of individuals who completed the training course at the training location, the impact of increasing quality from q_1 (first quartile) to q_4 (last quartile) is above 50 percent for all periods under analysis. Overall, these results confirm the strong causal effect of training quality on labor productivity independently of the program length.

Taken together, the estimates in Tables 4, 5, and 6 impart two related lessons. First, the on-the-job training experience matters in terms of overall treatment impacts. When comparing the impacts for those who completed only the first stage of the program (Row 1 in Table 6) with the overall impacts (Column 7 of Table 4), we observe large yet decreasing differences. Second, the returns to training quality are higher for the subsample of individuals who participate only in the first stage as compared to those for the whole sample, especially when looking at the marginal treatment impacts. This suggests that the on-the-job training experience has the ability to smooth the strong training quality effects on labor earnings across individuals attending low- and high-quality courses.

7.2 Ashenfelter's Dip and the Sustainability of the Treatment Effects

The difference-in-differences approach may be quite sensitive to the specific period over which the 'before" period is defined if we observe a drop in the mean earnings of participants prior to program entry (Ashenfelter 1978). Figure 3 depicts the earnings trajectory for the treatment and comparison groups. Three clear patterns emerge. Fist, there is some evidence about the existence of Ashenfelter's Dip in the PROJOVEN program. Because of data limitations, we cannot argument whether the pre-program drop in earnings is permanent or transitory. However, evidence from employment patterns in the months prior to the program is more consistent with the hypothesis of transitory drops in earnings, which imply that our estimates may be upward biased. Second, the pre-program earnings dip is similar for individuals attending both high- and low-quality courses. Thus, our marginal treatment impacts are not affected by the Ashenfelter's dip. Third, the full post-program earnings trajectory is very stable and consistent with the matching estimates emerging six, 12, and 18 months after the program. In particular, the increase of the quality premium across time is illustrative.

To avoid the sensitivity of our estimator to the pre-program earnings dip we implement cross-sectional matching estimators when the counterfactual is not participation. The results are somewhat similar with respect to those emerging from difference-in-difference models. For instance, the average treatment impacts for the pooled sample are 45, 37, and 28 percent six, 12, and 18 months after the program. For the same reference periods, the treatment impacts are 67, 50, and 50 for those attending high-quality courses and 36, 14, and 7 percent for those attending low-quality courses.

7.3 Repetition Effects

Because the training institutions can participate in successive bidding processes, they may improve their ability to present better proposals. Thus, the scores assigned to some quality proxies may be systematic correlated to "repetition" effects rather than to true quality improvements.[21] To avoid potential distortions in the evaluation of the proxies for quality that could upward bias the casual effect of training quality, we re-estimate the quality premiums by using only the subset of individuals attending training institutions that participate for first time in the program.[22]

Using the pooled data and identifying program impacts when the counterfactual is not participation, we find that the earnings gap between individuals attending high and low-quality courses (top and lowest quartiles) is 63, 51, and 35 percentage points after six, 12, and 18 months of program. These estimates are similar to the overall impacts presented in the last column of Table 4.

7.4 The Gender Dimension

Columns 9 and 10 of Table 2 show a potential damaging effect for the identification of the quality premiums: a disproportional number of males attend high-quality courses. Thus, the returns to training quality may not follow from gains in productivity but from intrinsic labor-market retribution to males' work. To purge this confounding factor from our estimates, we re-estimate the difference-in-differences estimator separately for males and females by using the pooled data.

Two basic patterns emerge from Table 7. First, the large overall treatment effects found in the PROJOVEN program is driven by the performance of the females participants. They show large and statistitically significant effects in both the short-run and the medium-term. On the other hand, the male participants show positive but smaller effects in the short-run and no effects in the medium-term. Second, the returns to training quality are positive and robust for all periods when looking the estimates emerging from the female subsample. We observe that the average treatment impacts for females attending high-quality courses are 96, 82, and 166 percent after six, 12, and 18 months of participation in the program. On the other hand, the treatment gains for males attending high-quality courses are positive but smaller (36, -8, and 12 percent, respectively).

In fact, the large number of female participants who moves from unproductive jobs (e.g., housekeepers and unpaid family workers) toward productive ones explains

[21] For instance, training institutions participating in successive bids can extract information to the program operator about the strengths an weaknesses of their prior proposals. This is particularly important for quality proxies such as curricular structure, contents, and activities whose evaluation is fuzzy.

[22] In average 9 new training institutions participated in each successive bid accounting for about one-third of the enrollment in each call. The quality index is re-estimated using only this subset of training institutions.

these striking estimates. In fact, 42 percent of female participants were working as either unpaid family workers or housekeepers before the program. Six months after the program, only 10 percent of them hold this type of jobs.

8. Conclusions

The adoption of market-based approaches in the provision of training services has shown to effectively increase the earnings of disadvantaged young individuals who frequently emerge from public schools operating in a far from any efficient frontier. For all cohorts, we find positive treatment impacts ranging in size from 20 to 60 percent 18 months after the program. These positive estimates are driven by the performance of female beneficiaries who show much larger treatment effects than male participants.

In fact, the combination of bidding processes with demand-driven approaches that ensures quality and pertinence of the training courses yield larger overall point estimates than those reported in the literature. These large gains are heightened because the PROJOVEN program targets relatively skilled individuals and the per-capita expenditures on participants are not small relative to the deficits that these programs are being asked to address. Thus, the PROJOVEN program has the ability to reallocate workers from unproductive jobs (e.g., unpaid family work, housekeeping) toward productive ones in firms protected by international laws that guarantee minimum work conditions.

We also find strong heterogeneity of the treatment gains across the quality dimension. Individuals attending high-quality training courses show much higher impacts than those attending low-quality courses. This result holds independently of the sample, matching method, and period of analysis. After controlling for treatment "dose" and pre-program earnings dip, we find that the quality premium is over 20 percentage points between individuals attending high- and low-quality courses.

This paper shows that the marginal returns to training quality are higher for those individuals who complete only the first stage of the program. This result indicates that the second stage of the program, the on-the-job training experience, smooth productive gains between people attending training courses of varying quality. The overall average treatment impacts, however, are higher for those who complete the paid, on-the-job training experience. From a cost-effectiveness perspective, this finding reveals the inadequacy of the program design, which mandates productive firms to assume the labor payments during the second stage of the program. It would be more effective if the program operator assumed the costs, given this is the key factor preventing almost half of the trainees from completing the second stage of the program.[23]

Finally, the reader should bear in mind that the strong quality premiums observed in this paper are based on a sample of training institutions that pass a minimum quality threshold imposed by the program operator. It is important to consider what the magnitude of these earnings differentials would be if training institutions located below the cut-off point were included.

[23] Because the law stipulates that trainees be paid the minimum wage during the on-the-job training, this cost is equivalent to around 700 soles (in real terms). Given the training "dose" premium is around 20 percentage points, and the average ex-post salary of those completing the whole program is around 300 soles, simple numerical calculation shows that in less than a year the benefits of completing the program surpass the salary costs of the on-the-job training experience.

References

Ashenfelter, O. 1978. “Estimating the Effect of Training on Earnings.” *Review of Economics and Statistics* 60: 47-57.

Behrman, J., Y. Cheng, and P. Todd. 2004. “Evaluating Pre-School Programs when Length of Exposure to the Program Varies: A Nonparametric Approach.” *Review of Economics and Statistics* 86(1): 108-132.

Behrman, J. and N. Birdsall. 1983. “The Quality of Schooling: Quantity Alone is Misleading.” *American Economic Review* 73(5): 928-946.

Bitler, M., J. Gelbach, and H. Hoynes. 2003. “What Mean Impacts Miss: Distributional Effects of Welfare Reform Experiments.” Cambridge, Mass.: National Bureau of Educational Research (NBER) Working Paper 10121.

Black, D. and J. Smith. 2003. “How Robust is the Evidence on the Effects of College Quality? Evidence from Matching.” *Journal of Econometrics* 121(1): 99-124.

-----. 2005. “Estimating the Returns to College Quality with Multiple Proxies for Quality.” Journal of Labor Economics, forthcoming..

Brewer, D., E. Eide, and R. Ehrenberg. 1999. “Does it Pay to Attend an Elite Private College? Cross-Cohort Evidence on the Effects of College Type on Earnings.” *Journal of Human Resources* 34(1): 104-123.

Campa, J. 1997. “Public Sector Retrenchment: Spain in the 1980s”, Department of Economics, New York University.

Card, D. and A. Krueger. 1992. “Does School Quality Matter? Returns to Education and the Characteristics of Public Schools in the United States.” *Journal of Political Economy* 100(1): 1-40.

Chacaltana, J. and D. Sulmont. 2003. “Políticas Activas en el Mercado Laboral Peruano: El Potencial de la Capacitación y los Servicios de Empleo.” Lima, Peru: Consorcio de Investigaciones Económicas y Sociales (CIES).

Dehejia R. and S. Wahba. 1999. “Causal effects in Non-Experimental Studies: Re-evaluating the Evaluation of Training Programs.” *Journal of the American Statistical Association* 94: 1053-1062.

Dale, S. and A. Krueger. 2002. “Estimating the Payoff to Attending a More Selective College: An Application of the Selection on Observables and Unobservables.” *Quarterly Journal of Economics* 117 (4): 1491-1528.

Fan, J. 1992. “Design-Adaptive Nonparametric Regression.” *Journal of the American Statistical Association* 87: 998-1004.

Fuller, B. 1987. “What School Factors Raise Achievement in the Third World.” *Review of Education Research* 57(3): 255-292.

Galdo, J. 1998. “La Evaluación de Proyectos de Inversión Social: Impacto del Programa de Capacitación Laboral Juvenil PROJOVEN.” *Boletín de Economía Laboral* 9, Ministerio de Trabajo y Promoción Social.

Galdo, J., D. Black, and J. Smith. 2006. “Bandwidth Selection and the Estimation of Treatment Effects with Non-Experimental Data.” Manuscript.

Glewwe, P. 2002. “Schools and Skills in Developing Countries: Education Policies and Socioeconomic Outcomes.” *Journal of Economic Literature* XL: 436-482.

Hanushek, E. 1986. “The Economics of Schooling: Production Function and Efficiency in Public Schools.” *Journal of Economic Literature* 24(3): 1141-1177.

Hanushek, E., Kain, J., O’Brien, D., and Rivkin, S. 2005. “The Market for Teacher Quality”, Manuscript.

Harbison, R. and E. Hanushek. 1992. *Educational Performance of the Poor: Lessons from Rural Northeast Brazil*. New York: Oxford University Press/World Bank.

Heckman, J. 2001. "Micro data, Heterogeneity, and the Evaluation of Public Policy: Nobel Lecture." *Journal of Political Economy* 109: 673-748.

Heckman J., and Pages C. 2000. "The Cost of Job Security Protection: Evidence from Latin America Labor Markets", NBER Working Paper #7773.

Heckman, J., H. Ichimura, J. Smith, and P. Todd. 1998. "Characterizing Selection Bias Using Experimental Data." *Econometrica* 66(5): 1017-1098.

Heckman, J., H. Ichimura, and P. Todd. 1997. "Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme." *Review of Economics Studies* 64(4): 605-654.

Heckman, J., H.R. LaLonde, and J. Smith. 1999. "The Economics and Econometrics of Active Labor Programs." In: O. Ashenfelter and D. Card, eds. *Handbook of Labor Economics, Volume 3A*. Amsterdam: North Holland, pp. 1865-2097.

Heckman, J., Layne-Ferrar, A., and Todd, P. 1995. "Does Measured School Quality Really Matter? An Examination of the Earnings Quality Relationship, NBER, Working Paper W5274

J. Heckman, R. Robb (1985): "Alternative Methods for Evaluating the Impact of Interventions: An Overview", *Journal of Econometrics*, 30 (1-2):239-267.

Ñopo, H., J. Saavedra, and M. Robles. 2001. "Una Medición del Impacto del Programa de Capacitación Laboral Juvenil PROJOVEN." Lima, Peru: Group for the Analysis of Development (GRADE).

Rama, M. 1999 "Efficient Public Sector Downsizing", *World Bank Economic Review*, 13, 1-22

Rosenbaum, P. and D. Rubin. 1983. "The Central Role of the Propensity Score In Observational Studies For Causal Effects." *Biometrika* 70(1): 41-55.

Smith, J. and P. Todd. 2005. "Does Matching Overcome LaLonde's Critique of Non-Experimental Estimators? *Journal of Econometrics* 125(1-2): 305-353.

Figure 1. Beneficiary Selection Process
Youth Training Program PROJOVEN, Lima 1996 to 2004.

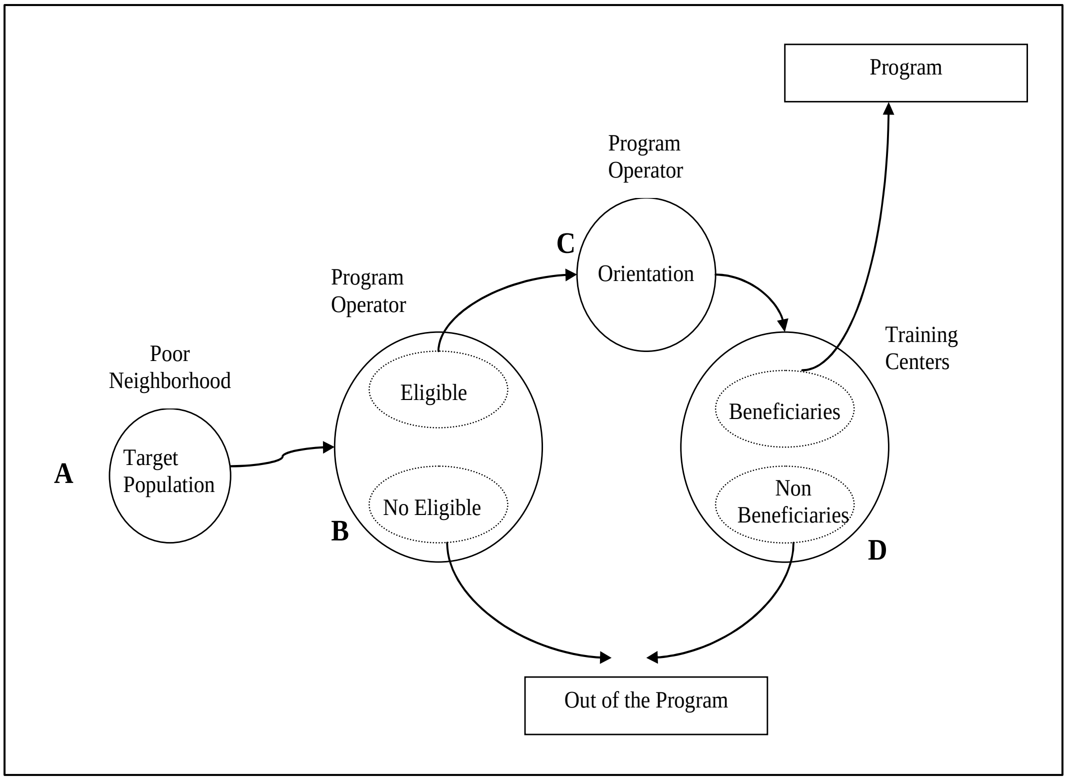
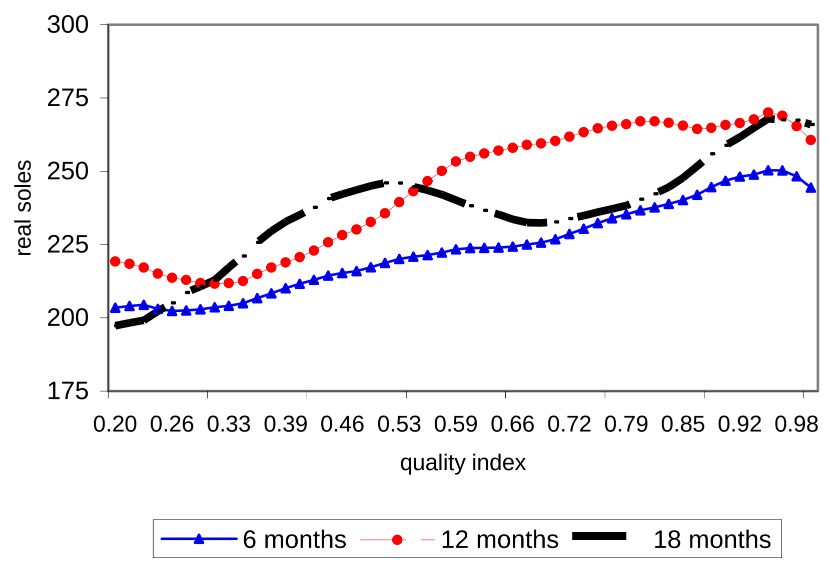


Figure 2: Local Linear Regression
PROJOVEN, Lima 1996-2004



Notes: Pooled data with bandwidth=0.20.

Figure 3: Monthly Earnings for Pooled Data
PROJOVEN, Lima 1996-2004

Note: Pooled means are unweighted. The quality index is constructed using the first principal component of factor analytic methods.

Table 1. Standardized Scores for Multiple Quality Proxies
Youth Training Program PROJOVEN, Lima 1996 to 2004

Quality Variables	Cohort 1	Cohort 2	Cohort 3	Cohort 4	Cohort 5
Class size	0.28	0.35	0.41	0.43	0.35
	(0.20)	(0.26)	(0.23)	(0.29)	(0.29)
Human resources	0.72	0.57	0.64	0.65	0.57
	(0.25)	(0.24)	(0.22)	(0.18)	(0.22)
Infrastructure	0.85	0.94	0.97	0.95	0.96
	(0.27)	(0.19)	(0.15)	(0.16)	(0.12)
Equipment	0.54	0.67	0.65	0.81	0.79
	(0.26)	(0.30)	(0.21)	(0.16)	(0.26)
Curricular structure	0.56	0.85	0.78	0.69	0.68
(contents and activities)	(0.28)	(0.26)	(0.24)	(0.29)	(0.27)
Market Knowledge	0.74	0.68	0.83	0.71	0.64
(worker and occupational analysis)	(0.22)	(0.21)	(0.15)	(0.19)	(0.21)
Expenditure per trainee	0.39	0.39	0.55	0.48	0.50
	(0.23)	(0.26)	(0.20)	(0.23)	(0.16)
PCA Quality Index	0.62	0.68	0.70	0.50	0.50
	(0.18)	(0.15)	(0.15)	(0.16)	(0.18)
# competing institutions	30	33	35	33	48
# competing courses	154	158	215	204	363
# funded courses	75	98	118	148	169

Notes: Each quality proxy is normalized as the ratio of the difference between the raw indicator value and the minimum value divided by the range. All normalized proxies are between 0 and 1. The quality index is constructed by principal component analysis based on first factors in all cases.

Table 2: Summary Statistics
PROJOVEN, Lima 1996-2004

	Pooled data		cohort 1	cohort 2	cohort 3	cohort 4	cohort 5	Quality Index	
	treated	comparison	p-value					treated >50th	treated <50th
A. Socio-Demographic									
age	19.67	19.73	0.02	0.51	0.22	0.92	0.84	19.69	19.64
sex (%)	42.70	42.60	0.93	0.97	0.91	0.88	0.94	46.00	40.00
schooling (%)									
none	0.23	0.25	0.08	0.31	0.97	0.31	----	0.12	0.34
incomplete primary	1.04	0.69	0.26	0.16	0.96	0.77	1.00	0.60	1.49
complete primary	4.82	6.27	0.38	0.22	0.29	0.46	0.85	4.64	5.05
incomplete high school	8.76	8.00	0.97	0.57	0.98	0.77	0.87	7.62	9.87
complete high school	85.14	84.70	0.53	0.92	0.88	0.79	1.00	87.02	83.24
marital status (%)									
single	91.19	77.44	0.00	0.00	0.00	0.00	0.00	92.02	90.38
married and/or cohabitating	5.12	14.87	0.00	0.00	0.00	0.00	0.00	5.24	5.27
other	3. 69	8.79	0.56	0.65	0.14	1.00	0.25	2.76	4.35
have children (%)	14.66	25.83	0.00	0.00	0.00	0.00	0.00	13.61	16.00
number of children	1.21	1.28	0.71	0.28	0.21	0.42	0.12	1.18	1.24
B. Labor information									
work status (%)									
have a job	52.17	52.11	0.54	0.84	0.88	0.94	0.94	54.52	50.00
unemployed	25.80	26.58	0.36	0.98	0.9	1.00	0.87	23.57	27.84
out of labor force	22.03	21.30	0.16	0.82	0.96	0.93	0.92	21.90	22.45
kind of work (%)									
self-employed	19.89	21.04	0.59	0.22	0.28	0.41	0.88	19.00	20.74
worker in private sector	53.67	62.23	0.00	0.71	0.98	0.42	0.03	55.46	51.18
worker in public sector	0.66	0.88	0.31	0.56	0.14	0.16	0.56	0.66	0.67
unpaid family worker	17.67	10.24	0.00	0.12	0.63	0.00	0.00	17.03	18.66
housekeeper	7.33	5.29	0.47	0.15	0.03	0.30	0.02	7.21	7.60
monthly earnings	91.54	126.00	0.00	0.02	0.08	0.00	0.00	99.34	83.40
experience	2.88	2.71	0.06	0.06	0.00	0.67	----	2.84	2.93
participation in training courses	23.03	23.00	0.37	0.32	0.02	0.03	0.00	24.70	21.50
hours of training	56.87	56.02	0.51	0.17	0.05	0.05	0.00	57.40	57.03
C. Household characteristics									
number of persons	6.23	6.00	0.00	0.44	0.00	0.10	0.28	6.26	6.22
household income	828.00	959.00	----	----	0.12	0.00	0.00	894.31	767.45
number of bedrooms	2.09	2.15	0.00	0.37	0.24	0.31	0.04	2.15	2.03
household density rate	3.12	2.87	0.05	0.90	0.00	0.72	0.00	3.10	3.13
floor: earthen	63.04	58.23	0.00	0.84	0.03	0.87	0.04	65.92	61.00
ceiling: concrete	35.07	23.98	0.00	0.29	0.00	0.00	0.00	34.13	35.85
walls: concrete/bricks	67.03	62.33	0.11	0.91	0.34	0.01	0.19	67.33	66.55
water: piped into the home	72.57	60.22	0.00	0.16	0.11	0.11	----	73.28	72.04
water sewage: flush toilet	65.95	61.28	0.00	0.33	0.15	0.01	0.10	65.87	65.95
D. Parent's schooling									
father (%)									
none	2.10	1.94	------	0.77	0.87	0.13	0.59	2.46	1.66
primary	37.52	32.78	------	0.63	0.35	0.35	0.00	37.77	37.76
incomplete high school	20.76	20.51	------	0.24	0.48	0.86	0.78	21.71	20.06
complete high school	26.72	32.22	------	0.77	0.94	0.00	0.00	26.77	26.83
higher education	7.57	5.20	------	0.29	0.37	0.01	0.29	6.37	8.85
mother (%)									
none	9.05	7.69	------	0.39	0.16	0.59	0.01	9.99	8.16
primary	47.27	41.93	------	0.89	0.04	0.00	0.34	47.90	47.16
incomplete high school	18.86	19.75	------	0.03	0.68	0.01	0.31	19.10	18.95
complete high school	18.09	21.90	------	0.82	0.69	0.00	0.07	16.79	19.36
higher education	3.72	3.19	------	0.41	0.97	0.00	0.80	3.47	4.01
#	1725	1742	599	627	720	732	764	840	873

Notes: Pooled means are unweighted. Not all means are based on the same number of observations because missing information for some individuals in the data. p-values refers to the test for differences in means for the treatment and comparison samples. Not all covariates are observed for all cohorts

Table 3: Coefficient Estimates from Balanced Probit Models for Program Participation
PROJOVEN, Lima 1996-2004

	Coefficients				
covariates	cohort 1	cohort 2	cohort 3	cohort 4	cohort 5
A. Socio-demographic					
constant	-1.49	0.02	6.67	1.82	-0.20
age	0.02	-0.03	-0.05	-0.04	0.06**
sex	-0.01	-0.06	-0.07	0.05	0.23**
schooling					
none	-----	-0.44	-0.52	----	----
incomplete primary	1.53**	----	-0.87	----	----
complete primary	----	-0.20	-0.65**	0.17	-1.09
incomplete high school	0.37	----	----	0.71	-0.88
complete high school	0.13	-0.13	-0.22	0.36	-1.11*
marital status					
single	0.26	0.62	-5.67**	-0.25	-0.56
married and/or cohabitating	-0.69	-0.06	-6.47**	-0.53	-1.12**
have children	-0.23	-0.02	-0.12	-0.35	0.38
number of children	0.05	-0.44	-0.01	-0.09	-0.68
B. Labor information					
work status					
have a job	-0.20	-0.51	-1.05**	-0.14	0.47
unemployed	-0.45**	-0.04	-0.19	0.01	-0.16
kind of work					
self-employed	0.38	1.23**	0.95**	0.62	-0.60
worker in private sector	0.18	1.04**	1.15**	0.63**	-0.88
worker in public sector	0.26	1.99*	----	----	-0.97
unpaid family worker	1.56**	0.74**	0.97**	0.72**	0.57
housekeeper	0.27	0.63	1.89**	0.92**	-0.15
monthly earnings	-0.00**	-0.00**	-0.00**	-0.00**	0.00
experience	0.03	0.34**	0.16**	0.10**	----
participation in training courses	-0.80**	0.34	-0.55**	0.20	0.25
hours of training	0.00**	-0.00**	0.00**	0.00	-0.00**
C. Household characteristics					
number of persons	0.06**	-0.05**	-0.25**	0.02	0.02
household income	----	----	----	-0.00*	-0.00**
number of rooms/ number of persons	0.06*	0.01	0.09**	0.00	0.16**
participation in welfare programs	0.14	----	----	----	----
floor : earthen	1.00**	0.31**	-0.11	-0.04	0.09
ceiling					
concrete	0.86**	0.01	0.36**	0.26**	-0.39
matting	----	-0.12	0.09	0.35**	-0.27**
walls: concrete /brick	-0.71**	0.15	-0.09	0.17	0.03
water: piped into the home	0.52**	0.81**	0.28**	----	----
water sewage: flush toilet	-0.24*	-0.55*	-0.38**	0.27**	0.28**
D. Parent's schooling					
father					
no information	----	----	0.63	-0.63	----
none	----	----	----	-0.25	-0.09
primary	----	-0.27	0.20	-0.42	0.31
incomplete high school	----	-0.02	-0.07	-0.36	0.24
complete high school	----	-0.17	0.00	-0.46*	0.09
higher education	----	-0.01	0.16	----	0.49
mother					
no information	----	----	-1.43**	----	----
none	----	0.70**	-0.30	-0.79**	0.74
primary	----	0.30	-0.18	-0.65*	0.33
incomplete high school	----	0.72**	0.00	-1.14**	0.17
complete high school	----	0.78**	-0.11	-1.19**	0.27
higher education	----	0.18	0.22	----	0.29
#	585	604	679	690	705
R^2	0.34	0.23	0.17	0.15	0.17

Note: * statistically significant at 5 percent, ** statistically significant at 10 percent. The specification of each probit model follows Dahejia and Wahba's (1999) balancing test. Not all covariates are observed for all cohorts.

Table 4. Average Treatment Impacts on Monthly Earnings
Difference-in-Differences Local Linear Matching Estimator
PROJOVEN, Lima 1996 to 2004

	Cohort 1	Cohort 2	Cohort 3	Cohort 4	Cohort 5	Pooled Data
A. 6 months after program						
$\Delta = Y_1(q) - Y_0(0)$	76 (35) [70]	58 (34) [52]	43 (24) [40]	42 (24) [33]	72 (27) [67]	58 [52]
$\Delta = Y_1(q_4) - Y_0(0)$	143 (49) [131]	89 (49) [79]	27 (45) [26]	22 (49) [17]	58 (41) [54]	65 [58]
$\Delta = Y_1(q_1) - Y_0(0)$	61 (53) [56]	35 (46) [31]	30 (54) [28]	36 (40) [28]	80 (41) [74]	49 [43]
B. 12 months after program						
$\Delta = Y_1(q) - Y_0(0)$	25 (34) [24]	82 (41) [73]	12 (32) [11]	-6 (26) [-5]	94 (32) [89]	43 [38]
$\Delta = Y_1(q_4) - Y_0(0)$	38 (60) [36]	125 (84) [110]	21 (48) [19]	-24 (44) [-19]	93 (35) [79]	51 [46]
$\Delta = Y_1(q_1) - Y_0(0)$	-17 (44) [-15]	75 (60) [67]	-29 (44) [-26]	-44 (41) [-35]	127 (48) [118]	27 [24]
C. 18 months after program						
$\Delta = Y_1(q) - Y_0(0)$	63 (35) [58]	68 (38) [61]	36 (32) [34]	25 (44) [20]	----	47 [42]
$\Delta = Y_1(q_4) - Y_0(0)$	108 (66) [100]	105 (64) [85]	101 (58) [94]	-17 (66) [-13]	----	71 [63]
$\Delta = Y_1(q_1) - Y_0(0)$	39 (70) [36]	75 (49) [67]	-3 (48) [-2]	2 (54) [1]	----	27 [24]

Notes: Point estimates are in real soles. Bootstrapped standard errors are in parenthesis. Percentage gains with respect to earnings in the baseline period are in brackets. Pooled data estimates are weighted averages from individual cohort estimates. The weights are the number of treatment units in each cohort. Difference-in-differences matching is applied to the sample of individuals inside the overlapping support region. q_4 and q_1 are the top and lowest quartiles of the quality index distribution.

Table 5. Marginal Treatment Impacts on Monthly Earnings
Difference-in-Differences Local Linear Matching Estimator
PROJOVEN, Lima 1996 to 2004

	Cohort 1	Cohort 2	Cohort 3	Cohort 4	Cohort 5	Pooled Data
A. 6 months after program						
$\Delta = Y_1(q_4) - Y_1(q_1)$	122 (63) [113]	51 (95) [46]	-29 (102) [-27]	29 (58) [23]	-38 (46) [-35]	22 [20]
$\Delta = Y_1(q_4) - Y_0(q_3)$	125 (60) [115]	6 (58) [5]	-110 (66) [-101]	17 (81) [20]	12 (51) [11]	10 [9]
B. 12 months after program						
$\Delta = Y_1(q_4) - Y_1(q_1)$	72 (55) [67]	5 (83) [4]	116 (80) [107]	46 (52) [36]	-54 (84) [-50]	17 [15]
$\Delta = Y_1(q_4) - Y_0(q_3)$	-33 (38) [-31]	43 (69) [38]	-25 (73) [-23]	11 (56) [9]	27 (70) [25]	6 [5]
C. 18 months after program						
$\Delta = Y_1(q_4) - Y_1(q_1)$	88 (63) [81]	14 (93) [12]	69 (88) [64]	12 (52) [9]	----	44 [39]
$\Delta = Y_1(q_4) - Y_0(q_3)$	118 (49) [109]	50 (76) [45]	11 (86) [10]	-1 (55) [-1]	----	42 [38]

Notes: Point estimates are in real soles. Bootstrapped standard errors are in parenthesis. Percentage gains with respect to earnings in the baseline period are in brackets. Pooled data estimates are weighted averages from individual cohort estimates. The weights are the number of treatment units in each cohort. Difference-in-differences matching is applied to the sample of individuals inside the overlapping support region. q_4, q_3, and q_1 are the fourth, third, and lowest quartiles of the quality index distribution.

Table 6. Treatment Impacts on Monthly Earnings for Formal Training
Difference-in-Differences Local Linear Matching Estimator
PROJOVEN, Lima 1996 to 2004

	Treatment Impacts for Pooled data		
	6 months after program	12 months after program	18 months after program
A. Average treatment impacts on the treated			
$\Delta = Y_1(q) - Y_0(0)$	21 (16) [19]	13 (12) [12]	31 (29) [28]
$\Delta = Y_1(q_4) - Y_0(0)$	55 (35) [49]	40 (37) [36]	59 (40) [53]
$\Delta = Y_1(q_1) - Y_0(0)$	3 (31) [3]	-14 (24) [-13]	45 (73) [40]
B. Marginal treatment impacts			
$\Delta = Y_1(q_4) - Y_1(q_1)$	53 (26) [47]	57 (32) [50]	52 (46) [47]
$\Delta = Y_1(q_4) - Y_0(q_3)$	36 (27) [33]	-6 (39) [-5]	82 (60) [73]

Notes: Pooled data estimates for the sub-sample of individuals that complete the training course at the training center location. Point estimates are in real soles. Bootstrapped standard errors are in parenthesis. Percentage gains with respect to earnings in the baseline period are in brackets. Difference-in-differences matching is applied to the sample of individuals inside the overlapping support region. q_4, q_3, and q_1 are the fourth, third, and lowest quartiles of the quality index.

Table 7. Treatment Impacts on Monthly Earnings for Males and Females Sub-samples
Difference-in-Differences Local Linear Matching Estimator
PROJOVEN, Lima 1996 to 2004

.

	Treatment Impacts for Pooled Data					
	6 months after the program		12 months after the program		18 months after the program	
	male	female	male	female	male	female
$\Delta = Y_1(q) - Y_0(0)$	37 (16) [33]	92 (21) [82]	2 (20) [2]	56 (31) [50]	1 (33) [1]	114 (62) [100]
$\Delta = Y_1(q_4) - Y_0(0)$	41 (32) [36]	108 (25) [96]	-9 (39) [-8]	92 (36) [82]	14 (47) [12]	187(60) [166]
$\Delta = Y_1(q_1) - Y_0(0)$	11 (36) [10]	91 (33) [81]	-12 (29) [-10]	35 (28) [31]	-35 (44) [-31]	88 (64) [78]

Notes: Pooled data estimates from males and females sub-samples. Point estimates are in real soles. Bootstrapped standard errors are in parenthesis. Percentage gains with respect to earnings in the baseline period are in brackets. Difference-in-differences matching is applied to the sample of individuals inside the overlapping support region. q_4 and q_1 are the top and lowest quartiles of the quality index distribution.

Appendix: Table A.1

Category	Variables
1. Class Size	• Number of trainees per course
2. Expenditures per capita	• Total cost per trainee
3. Human resources	• Number of instructors • Number of assistants • Teaching experience • Relationship between instructors' education and training course • Relationship between instructors' experience and training course • Instructors' training • Instructor's experience teaching disadvantage individuals
4. Infrastructure	• Proportion of classrooms • Quality of classrooms
5. Equipment	• Quantity of equipment, tools, and materials. • Quality of equipment, tools, and materials. • Availability of safety equipment
6. Curricular structure	• Analysis of contents • Relationship between contents and goals • Relationship between activities and goals • Relationship between activities and contents • Relationship between activities and methodology • Activities and safety regulations • Coherence between activities, equipment, and methodology. • Relationship between hours assigned to each activity and course goals. • Instruments of evaluation • Relationship between evaluation and goals • Frequency of evaluations • Pertinence of evaluation instruments for target population.
7. Market Knowledge	• Description of specific labor skills needed in that occupation • Relationship between activities and labor skills • Coherence between frequency of activities and labor skills • Inclusion of basic work competences (reading, speaking, interview skills) • Inclusion of equipment and tools needed in that occupation • Specific safety regulations in that occupation

Source: PROJOVEN's Bidding Process Instruments.