DOCUMENT RESUME

ED 457 359 CE 082 387

AUTHOR	Olsen, Robert B.; Decker, Paul T.
TITLE	Testing Different Methods of Estimating the Impacts of Worker Profiling and Reemployment Services Systems.
INSTITUTION	Mathematica Policy Research, Washington, DC.
SPONS AGENCY	Employment and Training Administration (DOL), Washington, DC. Office of Policy and Research.
PUB DATE	2001-06-00
NOTE	44p.; MPR Reference No. 8727-004.
CONTRACT	F-6828-8-00-80-30(07)
AVAILABLE FROM	For full text: http://wdr.doleta.gov/opr/fulltext/01-testwprss.pdf.
PUB TYPE	Reports - Evaluative (142)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	Adults; Comparative Analysis; Data Analysis; Evaluation Methods; Job Search Methods; *Matched Groups; Outcomes of Education; Profiles; *Program Effectiveness; Program Evaluation; *Regression (Statistics); *Reliability; *Research Methodology; Research Utilization; Retraining; *Unemployment Insurance
IDENTIFIERS	Florida

ABSTRACT

The Worker Profiling and Reemployment Services (WPRS) program requires states to establish systems for identifying Unemployment Insurance (UI) claimants likely to exhaust their UI benefits and refers them to reemployment services. An evaluation was conducted to assess the reliability of the impact estimates provided in the evaluation of the WPRS program, and to compute revised estimates of the impacts of WPRS programs if a more accurate estimation method could be identified. Data for the evaluation were gathered from the Job Search Assistance Demonstration in Florida, which, beginning in 1995, randomly assigned UI claimants to control groups or treatment groups who received training in job search techniques. The data were to be tested in two phases: In Phase I, using the regression method; and in Phase II, using variants of the matching methods used in other evaluations. The Phase I evaluation found that the linear regression model used in the WPRS evaluation produced accurate impact estimates, while the matched comparison groups tested in this evaluation produced less accurate impact estimates than the linear regression model. Based on the results of Phase I, therefore, it was decided not to proceed to Phase II of the evaluation. (Contains 10 references.) (KC)

Contract No.: F-6828-8-00-80-30 (07)
MPR Reference No.: 8727-004

Testing Different Methods of Estimating the Impacts of Worker Profiling and Reemployment Services Systems

June 2001

Robert B. Olsen
Paul T. Decker

Submitted to:

U.S. Department of Labor
Employment and Training Administration
200 Constitution Avenue, NW
Room N-5637
Washington, DC 20210

Project Officer:
Daniel Ryan

Submitted by:

Mathematica Policy Research, Inc.
600 Maryland Avenue, S.W.
Suite 550
Washington, DC 20024-2512
(202) 484-9220

Project Director:
Robert Olsen

This report was prepared for the U.S. Department of Labor, Employment and Training Administration, Office of Policy Research under contract F-6828-8-00-80-30 Task Order #7 with Mathematica Policy Research, Inc. Since contractors conducting research under government sponsorship are encouraged to express their own judgments freely, this report does not necessarily represent the office opinion or policy of the U.S. Department of Labor. The contractor is solely responsible for the contents of this report.

ACKNOWLEDGMENTS

The authors would like to thank a variety of individuals who have contributed to this evaluation and the completion of the analysis for this final report.

At the U.S. Department of Labor, Dan Ryan, our project officer, provided valuable guidance and assistance throughout the project; Steve Wandner provided valuable suggestions during the evaluation; and Wayne Gordon provided reports with important data from the Job Search Assistance Demonstration. At the Office of Workforce Innovation in Florida, Joan Hagan provided weekly data on the number of claimants assigned to WPRS in the local offices that participated in the Job Search Assistance Demonstration.

We received assistance at Mathematica from several individuals. Jon Jacobson reviewed and gave useful suggestions on the design memos and on this final report; Mary Grider provided excellent computer programming; and Donna Dorsey prepared the final document.

Robert Olsen
Project Director

CONTENTS

TABLES

FIGURES

EXECUTIVE SUMMARY

This evaluation is motivated by two goals: (1) to assess the reliability of the impact estimates provided in the evaluation of the Worker Profiling and Reemployment Services (WPRS) programs, and (2) to compute revised estimates of the impacts of WPRS programs if a more accurate estimation method can be identified. The evaluation also provides general information on the accuracy of different methods for estimating impacts without random assignment.

Under WPRS, states were required to establish systems for identifying Unemployment Insurance (UI) claimants likely to exhaust their UI benefits and referring them to reemployment services, such as resume preparation and training in job search methods. In an evaluation sponsored by the U.S. Department of Labor (USDOL), the impacts of WPRS were estimated by comparing UI claimants who were assigned to WPRS services (the treatment group) to claimants who were not assigned to WPRS services (the comparison group). Linear regression techniques were used to control for pre-existing differences between the two groups.

The results from the WPRS evaluation suggest that the impacts of WPRS on earnings are positive in some states and negative in others. However, the wide variation in impact estimates across states raises questions about the accuracy of the estimates. Furthermore, when the pre-existing differences between the treatment and comparison groups are large--as in the WPRS evaluation--linear regression methods can be unreliable. Therefore, the wide state-to-state variation in the estimated earnings impacts *may* be due to estimation error attributable to the regression method used in the WPRS evaluation.

Prior to the implementation of WPRS, USDOL sponsored a demonstration to test different program models that are consistent with the regulations governing WPRS. In 1995, the Job Search Assistance (JSA) Demonstration was implemented in the District of Columbia and in selected counties in Florida. Because the demonstration was based on the random assignment of eligible claimants to treatment and control groups, impacts were estimated by comparing treatment group members to control group members. Random assignment ensured that the pre-existing differences between the two groups were negligible.

Therefore, the demonstration should provide reliable estimates of the impacts of different WPRS program models via treatment-control differences. Furthermore, demonstration data can be used to compute other impact estimates using data that mimic the treatment and comparison samples available to the WPRS evaluation. The reliability of these impact estimates can be tested by comparing them to the treatment-control differences.

In this evaluation, we use data from the JSA Demonstration in Florida to mimic the treatment and comparison samples from the WPRS evaluation, and to test different methods of estimating impacts from these samples. These methods include the regression method used in the WPRS evaluation, but also include variants of the matching methods used in other evaluations. Matching is designed to select a subgroup of comparison group members who are similar to treatment group members. Impacts are then estimated by comparing treatment group members to the subgroup of similar comparison group members.

The plan for the evaluation included two phases:

- ***Phase I: Testing Different Methods of Estimating Impacts Using JSA Data.*** In Phase I, use data from the JSA Demonstration to assess the reliability of the regression method employed in the WPRS evaluation and the matching methods developed in this evaluation.

- ***Phase II: Applying Matching Methods to Actual WPRS Data.*** If any of the matching methods produce more accurate impact estimates than the regression method, apply those matching methods to WPRS data to obtain revised estimates of the impacts of WPRS on earnings.

DESIGN OF PHASE I OF THE EVALUATION

The design of Phase I consisted of two components: (1) identifying the analysis samples from JSA Demonstration data; and (2) specifying methods for estimating the impacts of being assigned to JSA/WPRS services on the claimants *who would have been assigned to services if WPRS had been operating in Florida in place of the demonstration.*

Identifying Three Samples from JSA Demonstration Data. We used the rule by which UI claimants are assigned to WPRS to determine which claimants would have been assigned to WPRS had it been operating instead of the demonstration. Claimants who would have been assigned to WPRS were classified as "treatment claimants" or "control claimants" for this evaluation based on their treatment-control status in the demonstration. Claimants who would *not* have been assigned to WPRS (and were not treated in the demonstration) were classified as "comparison claimants".

Specifying the Methods for Estimating Impacts. Based on the three analysis samples, we specified alternative methods of estimating the impacts of being assigned to WPRS. The experimental benchmark estimate equals the mean earnings of treatment claimants minus the mean earnings of control claimants. This benchmark is used to assess whether accurate impact estimates can be computed from "nonexperimental data"--data on treatment and *comparison* claimants--using either the linear regression method from the WPRS evaluation or one of the matched comparison groups developed for this evaluation.

The matching methods developed for this evaluation are designed to select "matched comparison groups" that look like the treatment group. A comparison claimant is selected for the matched comparison group if he or she can be "matched" to one or more treatment claimants with similar characteristics. The rules developed for defining acceptable matches require that matched claimants have the same sex, race/ethnicity, and education. Furthermore, matching claimants must have similar values for one of the following three variables:

1. ***Profiling Score.*** UI claimants are assigned "profiling scores" that reflect the probability of exhausting UI benefits without additional reemployment services, and are assigned to WPRS based on these scores.

2. ***Base-year Earnings.*** Claimants are determined eligible for UI based on their "base-year earnings", which measures total earnings in four out of five quarters prior to the UI claim.

3. ***Propensity Score.*** Treatment claimants have higher probabilities or propensities of being assigned to services than comparison claimants, and "propensity scores" are often computed in evaluations to use as matching variables.

FINDINGS FROM PHASE I OF THE EVALUATION

Based on the treatment and control groups in this evaluation, the experimental benchmark estimate that we use to assess the accuracy of other impact estimates equals $260. Therefore, the average earnings of treatment claimants in the year following the quarter of random assignment were $260 higher than the average earnings of control claimants in the same year.

How well did the different methods for estimating earnings impacts from the treatment and comparison samples perform? The two main findings from Phase I of the evaluation are given below:

1. ***The linear regression model used in the WPRS evaluation produced accurate impact estimates.*** The estimate produced by the linear regression model from the WPRS evaluation equals $308, which is very close to the experimental benchmark of $260.

2. ***The matched comparison groups tested in this evaluation produced less accurate impact estimates than the linear regression model.*** The impact estimates based on matched comparison groups range from -$111 to -$3,440, and none of these estimates are as close to the experimental benchmark as the estimate produced by the linear regression model.

Therefore, despite the general concerns that can be raised about the reliability of regression methods to adjust for large differences between treatment and comparison groups, *this evaluation provides no evidence that the regression methods used in the WPRS evaluation are unreliable.*

The poor performance of the matching methods tested in this evaluation can be attributed to the difficulty in selecting matched comparison groups that are sufficiently similar to the treatment group. Each matched comparison group was similar to the treatment group on many dimensions but different from the treatment group in at least one dimension that proved to be important. None of the matched comparison groups had the same (or a very similar) distribution of claimants across the local offices in the demonstration as the treatment group. Findings in this report suggest that it may be impossible to create a matched comparison group that is comparable to the treatment group in the distribution of claimants across local offices, and is *also* comparable to the treatment group in other important dimensions, such as sex, race/ethnicity, education, the profiling score, base-year earnings, and the propensity score. In other words, we were unable to create a matched comparison group that was comparable to the treatment group on all the dimensions that seemed important.

REFERENCES

Cochran, W.G. (1965), "The Planning of Observational Studies in Human Populations," *The Journal of the Royal Statistical Society*, 128, 234-265.

Decker, Paul T., Irma L. Perez-Johnson, and Walter S. Corson (1997). "The Job Search Assistance Demonstration Implementation Report", Washington, DC: Mathematica Policy Research, Report to U.S. Department of Labor, Employment and Training Administration.

Decker, Paul T., Robert B. Olsen, Lance Freeman, and Daniel H. Klepinger (2000). "Assisting Unemployment Insurance Claimants: The Long-Term Impacts of the Job Search Assistance Demonstration" (OWS Occasional Paper 2000-02), Washington, DC: U.S. Department of Labor, Employment and Training Administration, Office of Workforce Security, http://wdr.doleta.gov/owsdrr/00-2/.

Dehejia, Rajeev, and Sadek Wahba (1999), "Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs," *Journal of the American Statistical Association*, 94, 1053-1062.

Dickinson, Katherine P., Suzanne D. Kreutzer, Richard W. West, and Paul T. Decker (1999), "Evaluation of Worker Profiling and Reemployment Services Systems," Report prepared for the U.S. Department of Labor, Employment and Training Administration, Menlo Park, CA: Social Policy Research Associates.

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

Heckman, James J., Hidehiko Ichimura, and Petra Todd (1998), "Matching as an Econometric Evaluation Estimator," *Review of Economic Studies*, 65, 261-294.

Rosenbaum, Paul R. (1995), *Observational Studies*, New York, NY: Springer-Verlag.

Rosenbaum, Paul R., and Donald B. Rubin (1983), "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika*, 70, 41-55.

Rosenbaum, Paul R., and Donald B. Rubin (1985), "Constructing a Control Group Using Multivariate Matched Sampling Methods That Incorporate the Propensity Score," *The American Statistician*, 39, 33-38.

APPENDIX A: WEIGHTING THE FOUR MATCHED COMPARISON GROUPS

The four matched comparison groups we selected were used to compute four different estimates of the impacts of the treatment on claimants who would have been assigned to WPRS. These impact estimates are computed by subtracting the average earnings of matched comparison group members from the average earnings of treatment group members. However, these averages are not simple, unweighted averages. As shown previously in Table 4, we assign weights to treatment claimants that reflect the sampling probabilities in the local offices where they applied for benefits. As described in this section, we assign weights to matched comparison group members that reflect the sampling probabilities *of the treatment group members to which they were matched*. The weights for matched comparison claimants were designed to ensure that the sum equals the sum of the weights for treatment claimants.

Weighting the treatment group was straightforward:

- Treatment claimants matching at least one comparison claimant are assigned weights based on Table 4.
- Unmatched treatment claimants are assigned weights of zero, which effectively dropped them from the analysis.[15]

Dropping unmatched treatment claimants from the analysis is undesirable because if too many treatment claimants are dropped, it becomes difficult to generalize the estimated impacts for matched treatment claimants to all treatment claimants. However, including unmatched treatment claimants in the analysis would guarantee that the treatment and matched comparison groups are systematically different due to the unmatched treatment claimants, and would therefore defeat the point of selecting matched comparison groups.

Each matched comparison claimant is weighted to reflect the number of claimants represented by the treatment claimants to whom he or she is matched. Suppose, for example, a comparison claimant is matched to two treatment claimants--one from Clearwater and the other from Davie. Table 4 indicates that the weights for these treatment claimants are 3.76 and 7.44, respectively. Under matching rules 1-3, the weight assigned to the matched comparison claimant would be 11.20, the sum of the two treatment weights, to account for the 3.76 claimants represented by each treatment claimant in Clearwater and the 7.44 claimants represented by each treatment claimant in Davie.

Weighting the matched comparison group generated by matching rule 4 is more complicated because this rule allows each treatment claimant to match multiple comparison claimants. The multiple comparison claimants matching a single treatment claimant are weighted so that *together* they reflect the claimants represented by the treatment claimant. Consider the example from the previous paragraph, but suppose that a second comparison claimant matched the

[15] Unmatchable treatment claimants are *not* dropped from the analysis files used to compute the regression-based estimates. However, regression-based estimates can be heavily influenced by outliers, such as peculiar treatment units that cannot be matched to any comparison units. Furthermore, regression-based estimates will often provide poor impact estimates for treatment units without any similar comparison units.

treatment claimant from Davie. We would weight each of the two matching comparison claimants to represent half of the claimants represented by the treatment claimant from Davie, or 3.72 claimants (half the treatment group weight of 7.44). However, the first of two comparison claimants was also matched to the treatment claimant from Clearwater, so this comparison claimant would receive a weight of 7.48 (3.72 + 3.76).

The weights for matched comparison claimants depend on the number of treatment claimants to which they matched. Furthermore, different random samples based on the sampling plan described in Chapter II would produce different treatment and comparison samples, different matched pairs of treatment and comparison claimants, and different weights for matched comparison claimants. Therefore, the weights for matched comparison claimants contain sampling variation. When weights contain sampling variation, the least squares estimate of the standard error of the impact estimate will be biased. Therefore, it is necessary to account for sampling variation in the weights when estimating the standard errors of the impact estimates generated by our four matched comparison groups. Our solution to this problem is to estimate these four standard errors via bootstrapping. For comparability between the standard errors of all six impact estimates presented in this report, we estimate the standard errors of all six impact estimates via bootstrapping.

Based on the results from Phase I, MPR recommended not to proceed to Phase II of this evaluation, and USDOL concurred. This recommendation was based on the finding that the regression methods used in the WPRS evaluation produced accurate estimates of earnings impacts from the demonstration data. This report provides no evidence that matched comparison groups of the types we tested would yield more accurate estimates of the impacts of WPRS.

CHAPTER I: INTRODUCTION

This evaluation is motivated by two specific goals and one more general goal. The specific goals are to (1) assess the reliability of the impact estimates provided in the evaluation of the Worker Profiling and Reemployment Services (WPRS) programs, and (2) compute revised estimates of the impacts of WPRS if a more accurate estimation method can be identified. The more general goal of this evaluation is to provide information about which estimation methods are most accurate when computing impact estimates from nonexperimental data--data without a randomly assigned control group.

WPRS was created in response to a 1993 amendment to the Social Security Act. This amendment required states to establish profiling systems for targeting Unemployment Insurance (UI) claimants likely to remain unemployed long enough to exhaust their UI benefits, and for referring targeted claimants to reemployment services shortly after they apply for benefits (Dickinson et al., 1999). The program model implemented under WPRS varies across states. In some states, all claimants assigned to the program are required to participate in the same set of services. In other states, counselors have more discretion in specifying the services in which each claimant must participate to remain eligible for UI benefits. These services include training in job search methods, resume preparation, job development, and referrals to job openings.

Prior to the implementation of WPRS, the U.S. Department of Labor (USDOL) sponsored a demonstration to test different program models that are consistent with the regulations governing WPRS. The Job Search Assistance (JSA) Demonstration was implemented in the District of Columbia and selected counties in Florida in 1995, and it continued to operate in 1996 when the implementation of WPRS began. The evaluation of the demonstration (Decker et al., 2000) was based on the random assignment of eligible claimants to three treatment groups and one control group. The first treatment was based on a list of services in which all claimants were required to participate. The second two treatments allowed counselors to determine the required services for each claimant on an individualized basis. The key outcome variables in this evaluation were (1) UI benefits and duration, and (2) employment and earnings. Because the treatment impacts were measured *relative to a randomly selected control group*, the resulting impact estimates are more credible than those measured relative to a nonrandom comparison group whose members may differ systematically from treatment group members.

During the evaluation of the JSA Demonstration, USDOL also sponsored an evaluation of the WPRS program itself in six states (Dickinson et al., 1999). The WPRS evaluation can be justified by two problems with generalizing the findings from the demonstration to WPRS. First, the rule by which UI claimants were assigned to demonstration services was different from the rule by which UI claimants are assigned to WPRS services in most states. Therefore, WPRS targets a different set of claimants than those who would have been eligible for the demonstration. Second, WPRS service models differ across counties and states, so the results from the District of Columbia and ten counties in Florida may be unrepresentative of WPRS nationwide.

Unlike the JSA Demonstration, the WPRS evaluation lacked the benefit of a randomly assigned control group: impacts were estimated by comparing UI claimants assigned to WPRS services (the "treatment" group) to a comparison group consisting of claimants who were not assigned to WPRS services. Due to the rule by which claimants are assigned to services, comparison group members in the evaluation had systematically different baseline characteristics than treatment group members. The WPRS evaluation accounted for baseline differences between the two groups using regression analysis. However, when the differences between two groups are large, linear regression can produce biased impact estimates (Cochran, 1965). Furthermore, the estimated earnings impacts from the WPRS report varied considerably across states (Dickinson et al., 1999, Exhibit III-9.). This variation across states may be "real": some state programs may be much more effective than others. However, given the difficulties in estimating impacts without a randomly assigned control group, the wide variation in impact estimates across states raises questions about the accuracy of the estimates.

One alternative method of adjusting for baseline differences between treatment and comparison groups is "statistical matching". Each treatment group member is matched to one or more comparison group members with similar baseline characteristics. Comparison group members who are matched to one or more treatment group members are included in the "matched comparison group": other comparison group members are excluded. Matching is designed to select a subgroup of the comparison group that has similar baseline characteristics to the treatment group. The goal of matching is to select matched comparison group members whose outcomes are as similar as possible to what the outcomes of treatment group members would have been in the absence of the treatment.

One particular form of matching that has become increasingly popular is called propensity score matching (Rosenbaum and Rubin 1983, 1985). Among people with the same probability of participating in (or being assigned to) a program, whether or not a person actually participates is a purely random event like assignment to the treatment or control groups in a random-assignment experiment. Propensity score matching selects matched groups of treatment and comparison units with similar participation probabilities or "propensities"--or more typically, similar estimated propensities. Through propensity score matching, baseline differences between the treatment and comparison groups can be reduced for many baseline variables while using a single variable for matching. Dehejia and Wahba (1999) use experimental data from the National Supported Work Demonstration to show that propensity score matching can produce impact estimates that are very close to the experimental estimates. These results suggest that propensity score matching can generate accurate estimates of the impacts of some employment-related programs.

The results from Dehejia and Wahba raise the following question: *can propensity score matching generate accurate estimates of the impacts of WPRS*? Phase I of this evaluation was designed to test the reliability of three different matching methods, including propensity score matching, and the reliability of the regression methods used in the WPRS evaluation. Phase II of the evaluation was designed to apply matching methods to data from the WPRS evaluation to compute revised estimates of WPRS's impacts on earnings. However, Phase II would only proceed if Phase I showed that the matching methods produced more accurate impact estimates than the regression methods.

Therefore, this evaluation can be summarized as follows:

- ***Phase I: Testing Different Methods of Estimating Impacts with JSA Data.*** In Phase I, we used data from the JSA Demonstration to assess the reliability of the standard regression methods employed in the WPRS evaluation and the matching methods developed in this evaluation.

- ***Phase II: Applying Matching Methods to Actual WPRS Data.*** If any of the matching methods had produced more accurate impact estimates than the regression methods, we would have applied them to obtain revised estimates of the impacts of WPRS.

Based on the results from Phase I, MPR recommended not to proceed to Phase II of this evaluation, and USDOL concurred. This recommendation was based on the finding that the regression methods used in the WPRS evaluation produced accurate estimates of earnings impacts from the demonstration data. This report provides no evidence that matched comparison groups of the types we tested would yield more accurate estimates of the impacts of WPRS.

The remaining chapters of this report describe the evaluation's design and the results from Phase I. Chapter II describes the design of the analysis samples used in this evaluation. Chapter III describes the different methods we tested for estimating impacts. Chapter IV provides the impact estimates generated by applying different estimation methods to the analysis samples. These estimates support MPR's recommendation and USDOL's decision to end the evaluation after Phase I.

CHAPTER II: DESIGN OF THE ANALYSIS SAMPLES

The design of this evaluation required four steps: selecting the data, identifying analysis samples within the data, weighting the samples to ensure that the analysis samples are representative of the populations of interest, and specifying methods for estimating program impacts. As described in the introduction, we selected data from the JSA Demonstration because the demonstration's treatments roughly correspond to the types of service packages to which claimants are assigned under WPRS, and because the demonstration included a randomly assigned control group. This chapter describes the second and third design steps--identifying and weighting the analysis samples--and leaves the estimation methods to Chapter III.

A. IDENTIFYING THE ANALYSIS SAMPLES

The data used in this evaluation were collected to support the evaluation of the Job Search Assistance Demonstration in 10 local offices in Florida (Decker et al., 2000). The evaluation of the demonstration was designed to provide estimates of the impacts of three different job search assistance treatments *on claimants who were eligible for the demonstration*. From the claimants who applied for UI benefits in the 10 demonstration offices during the demonstration, the state of Florida identified claimants who were newly unemployed, who did not have a specific date of recall to their previous employer, who did not obtain jobs through a union hiring hall, and who met a few other eligibility criteria.[1] Final eligibility for the demonstration was determined based on profiling scores that were assigned to all claimants meeting the state's eligibility criteria. This score provided an estimate of the probability that the claimant would exhaust his or her entitlement of UI benefits in the absence of additional reemployment services.[2] Profiling scores were assigned based on a linear model of benefit exhaustion, which was estimated from historical UI data. The model of benefit exhaustion included the following variables as predictors: education, industry, occupation, and job tenure.[3] Claimants were deemed eligible for the demonstration if they met the state's eligibility criteria, and if they were assigned profiling scores greater than 0.4.

Under WPRS in Florida, claimants are assigned profiling scores using the model of benefit exhaustion developed for the demonstration. However, WPRS differs from the demonstration in how it uses these scores to target reemployment services. Under WPRS, the claimants with the highest profiling scores are assigned to locally provided services subject to local capacity constraints. Therefore, the average profiling score should be higher for claimants who would have been assigned to WPRS than for demonstration participants. If claimants with higher profiling scores have a greater need for reemployment services than other claimants, the average impacts of reemployment services might be higher for "WPRS-targeted" claimants than for demonstration participants.

[1] For a complete list of the state's eligibility criteria, see Decker et al., 1997, p. 40-41.

[2] Individuals who qualify for UI are entitled to a fixed amount of UI benefits, and most of those who remain on UI for 26 weeks exhaust their entitlement.

[3] The coefficient estimates from the profiling score model are provided in Table III.1 of Decker et al., 1997.

To develop a fair test of how different methods of estimating impacts would perform if applied to data from the WPRS evaluation, *we measure the impacts of the demonstration's three treatments on claimants who would have been assigned to WPRS.* To identify these claimants, we apply a stylized version of the process that local offices use under WPRS to assign profiled claimants to services.

The number of claimants assigned to WPRS in a particular local office and week depends on the capacity of the local office to provide reemployment services. It is not possible to determine with certainty the capacity of each local office in each week that the demonstration was operating. In fact, it is surprisingly difficult to determine *current* office capacity. However, given that we would like to generalize the results from this test to other counties and states with different local capacities, it is not critical that we use exact estimates of local office capacity in the demonstration offices. We distinguish between large, medium-sized and small offices using estimates of the average weekly number of claimants who applied for UI in each office and were "profiled" (assigned a profiling score). Based on these estimates, we define large offices to be those that profiled approximately 200 or more claimants per week (Davie, Ft. Lauderdale, Hialeah, and Orlando), medium-sized offices as those that profiled approximately 100 claimants per week (Clearwater and Miami), and small offices as those that profiled 25-75 claimants per week (Ft. Pierce, Lakeland, Pensacola, and St. Augustine). Also, we assume that larger offices have larger capacities to provide reemployment services than smaller offices, and that no office can serve more claimants than could be taught in a classroom setting once per week.[4] Based on these assumptions, we assigned office capacities of 30, 20, and 10 to large offices, medium-sized offices and small offices, respectively, as shown in Table 1. The numbers in Table 1 suggest that larger offices have larger capacities than smaller offices, but not *proportionately* larger capacities: larger offices serve a smaller proportion of profiled claimants than smaller offices.[5]

As a percentage of all profiled claimants, the capacity assumptions in Table 1 are well within the national distribution (Dickinson et al., Exhibit II-5). Furthermore, the estimates in Table A.3 are also roughly consistent with the number of claimants assigned to WPRS in the week of November 6, 2000—a week for which we were able to obtain data from Florida's Office of Workforce Innovation—in the six demonstration offices that appear to be operating a WPRS program. Lastly, there is no reason to believe that our impact estimates will be very sensitive to our assumptions about office capacity.

[4] These assumptions are consistent with observations from visits to local sites and with current data on the number of claimants assigned to WPRS in the demonstration offices.

[5] This is consistent with observations made during visits to local offices for the WPRS evaluation.

TABLE 1: ESTIMATED WEEKLY CAPACITY FOR FLORIDA LOCAL OFFICES UNDER WPRS

Florida Local Office	Average Number of Profiled Claimants	WPRS Capacity (Claimants)
Clearwater	106	20
Davie	193	30
Ft. Lauderdale	345	30
Ft. Pierce	52	10
Hialeah	267	30
Lakeland	52	10
Miami	107	20
Orlando	227	30
Pensacola	75	10
St. Augustine	34	10

With estimates of local office capacity, a complete list of profiled claimants with their profiling scores in a particular office and week would reveal the claimants who would have been assigned to WPRS--those with the X highest profiling scores, where X equals the capacity of the local office. The profiling score of the X^{th} claimant, P, defines the "WPRS threshold" in that office and week because claimants with profiling scores lower than P would not be assigned to WPRS in that office and week. However, the demonstration data do not include a complete list of profiled claimants. Some eligible claimants were excluded from the demonstration and therefore from the demonstration data because the demonstration was designed to accommodate 24 claimants per office and week (six for the control group and six for each of three treatment groups). Additional claimants who were eligible for the demonstration were excluded.

The exclusion of eligible claimants from the demonstration data complicates the process of identifying the claimants who would have been assigned to WPRS. Many excluded claimants had profiling scores that were high enough that they would have been assigned to WPRS. For example, if the capacity of a local office is 10 claimants per week, we cannot assume that the 10 claimants with the highest profiling scores in that office-week participated in the demonstration and can therefore be found in our demonstration data.[6]

Fortunately, eligible claimants were excluded from the demonstration on a random basis, so the demonstration participants are a simple random sample of the demonstration-eligible claimants in each office and week. Therefore, we assign a "demonstration weight" to each demonstration-eligible claimant who was randomly selected to participate in the demonstration. This weight equals the inverse of the selection probability. For example, suppose that one eligible claimant was excluded from the demonstration for each control group member. Since three claimants were assigned to the treatment group for each claimant assigned to the control group, the probability of being included in the demonstration--as either a treatment or control

[6] Assuming that the 10 top-ranked demonstration participants would be targeted for WPRS services violates the office's capacity constraint: some claimants with profiling scores that were higher than the profiling score of the 10th-ranked demonstration participant were excluded from the demonstration.

group member--equals 4/5^{ths} or 80 percent. The demonstration weight for treatment and control group members in this office equals the inverse of the selection probability or 1.25: *each demonstration participant (treatment or control group member) represents 1.25 eligible claimants in the same office.* Therefore, if a local office's capacity is 10 claimants per week, the 8 demonstration participants with the highest profiling scores represent the 10 demonstration-eligible claimants who would have been assigned to services. The demonstration weights for demonstration participants in each local office are provided in Table 2.

TABLE 2: DEMONSTRATION WEIGHTS FOR DEMONSTRATION PARTICIPANTS

Local Office	Selection Probability	Demonstration Weight
Clearwater	35%	2.82
Davie	18%	5.58
Ft. Lauderdale	10%	10.42
Ft. Pierce	92%	1.09
Hialeah	11%	9.35
Lakeland	89%	1.13
Miami	13%	7.85
Orlando	16%	6.14
Pensacola	80%	1.26
St. Augustine	86%	1.16

The weights in Table 2 were assigned to demonstration participants and combined with the capacity estimates in Table 1 to predict which claimants would have been assigned to WPRS. Table 3 provides an example of how to identify these claimants from the demonstration data in a single office and week. The first column of Table 3 shows the profiling score of each hypothetical claimant profiled in the Pensacola local office in a single week. Claimants are ordered from highest to lowest profiling score, and the second column provides the rank of each claimant. The third column indicates the random assignment of eligible claimants (those with profiling scores > 0.4) to four demonstration groups--the control group (*C)* the three treatment groups (*T1*, *T2*, and *T3*)--and one "nonexperimental" group (*N*) that did not participate in the demonstration. The third column also identifies demonstration-ineligible claimants with profiling scores below 0.4. The fourth column provides the demonstration weight of demonstration participants (rounded from 1.26 to 1.25 to simplify the example). The fifth column of Table 3 identifies members of each of the three analysis samples:

1. ***Treatment Group.*** Claimants who were (1) assigned to one of the demonstration's treatment groups, and (2) would have been assigned to WPRS.
2. ***Control Group.*** Claimants who were (1) assigned to one of the demonstration's treatment groups, and (2) would have been assigned to WPRS.
3. ***Comparison Group.*** Claimants who were (1) not assigned to one of the demonstration's treatment groups, and (2) would not have been assigned to WPRS.

The definitions of the three analysis samples describe how to identify sample members from the example in Table 3. Table 1 shows that our estimate of Pensacola's capacity to provide WPRS services is 10 claimants per week. Therefore, the 10 claimants with the highest profiling scores--shaded in gray in Table 3--would have been assigned to WPRS in Pensacola in this hypothetical week. However; two of these 10 claimants were assigned to the nonexperimental group and therefore cannot be found in the demonstration data. As described earlier, we adjust for the absence of these claimants from the demonstration data by assigning a weight of 1.25 to each demonstration participant. The eight demonstration participants with the highest profiling scores would have been assigned to WPRS because they represent the 10 demonstration-eligible claimants with the highest profiling scores. The treatment group for this evaluation consists of the six treatment group members who would have been assigned to WPRS; the control group for this evaluation consists of the two control group members who would have been assigned to WPRS. The comparison group consists of all untreated demonstration participants and ineligible claimants who would not have been assigned to WPRS because their profiling scores were too low. The example in Table 3 refers to a single local office in a single week, but the three analysis samples for this evaluation consist of claimants from all 10 demonstration offices and all 53 weeks of the demonstration.

B. WEIGHTING THE ANALYSIS SAMPLES

The three analysis samples are weighted to reflect sampling probabilities. The treatment and control groups are stratified random samples of the claimants who would have been assigned to WPRS. Consider a local office with capacity *X* (from Table 1). We compute the probability of being assigned to the control group, *P(c)*, and the probability of being assigned to the treatment group, *P(t)*, based on the probabilities in Table 2 and the treatment-control ratio of 3:1. These sampling probabilities are applicable to both the population of demonstration-eligible claimants in a local office *and* the subgroup of the population who would have been assigned to WPRS. We assign weights of 1 / *P(c)* to each control group member and 1 / *P(t)* = 1 / [3· *P(c)*] to each treatment group member in the local office. Weights for both the treatment and control groups are provided in Table 4.

The comparison sample is a stratified random sample of claimants who would not have been assigned to WPRS. This sample consists of two subgroups: (1) claimants from the demonstration's control group with profiling scores below the WPRS threshold, and (2) demonstration-ineligible claimants with profiling scores below 0.4. The first subgroup receives the same weight as all other control group members from the same office (see Table 4). The second subgroup is assigned weights that reflect the sampling rates for ineligible cases in the demonstration. Administrative data were collected for a stratified random sample of ineligible claimants, where the strata were based on the month of random assignment. We assigned analysis weights to demonstration-ineligible claimants that equal the inverse of the sampling probabilities, and these weights are displayed in Table 5.

TABLE 3: PROFILED CLAIMANTS IN ONE WEEK, PENSACOLA OFFICE
(Capacity = 10 Claimants)

PROFILING SCORE	RANK	DEMONSTRATION	WEIGHT	WPRS SAMPLE
.85	1	T1	1.25	Treatment
.84	2	C	1.25	Control
.82	3	N	0	
.80	4	T3	1.25	Treatment
.78	5	T2	1.25	Treatment
.77	6	T3	1.25	Treatment
.76	7	N	0	
.76	8	T1	1.25	Treatment
.73	9	T2	1.25	Treatment
.72	10	C	1.25	Control
.71	11	T1	1.25	
.71	12	C	1.25	Comparison
.70	13	T3	1.25	
.69	14	N	0	
.67	15	T2	1.25	
.67	16	N	0	
.66	17	C	1.25	Comparison
.65	18	T3	1.25	
.65	19	T2	1.25	
.63	20	T1	1.25	
.62	21	T1	1.25	
.59	22	C	1.25	Comparison
.58	23	T3	1.25	
.56	24	N	0	
.51	25	T2	1.25	
.50	26	N	0	
.48	27	C	1.25	Comparison
.46	28	T3	1.25	
.43	29	T2	1.25	
.41	30	T1	1.25	
.40	ELIGIBILITY THRESHOLD FOR THE DEMONSTRATION			
.39	31	Ineligible	N.A.	Comparison
.37	32	Ineligible	N.A.	Comparison
.33	33	Ineligible	N.A.	Comparison
.30	34	Ineligible	N.A.	Comparison
.25	35	Ineligible	N.A.	Comparison

TABLE 4: ANALYSIS WEIGHTS FOR TWO EXPERIMENTAL GROUPS

Local Office	Treatment Group	Control Group
Clearwater	3.76	11.37
Davie	7.44	22.31
Ft. Lauderdale	13.89	41.68
Ft. Pierce	1.45	4.34
Hialeah	12.47	37.42
Lakeland	1.50	4.51
Miami	10.47	31.40
Orlando	8.19	24.56
Pensacola	1.68	5.03
St. Augustine	1.55	4.65

TABLE 5: ANALYSIS WEIGHTS FOR COMPARISON GROUP (INELIGIBLES ONLY)

Month of Random Assignment	Sampling Probability	Weight
July 1995	12%	8.43
August 1995	13%	7.94
September 1995	16%	6.21
October 1995	15%	6.66
November 1995	16%	6.27
December 1995	18%	5.50
January 1996	20%	5.11
February 1996	17%	5.89
March 1996	21%	4.86

CHAPTER III: METHODS FOR ESTIMATING EARNINGS IMPACTS

This evaluation is designed to test different methods of estimating the impacts of WPRS with the type of nonexperimental data that were available to the WPRS evaluation. The previous chapter describes how we use data from the JSA Demonstration to simulate the treatment and comparison samples from the WPRS evaluation, and the control sample that is the key to this evaluation. In this chapter, we describe how we use the treatment and control samples to compute the experimental benchmark estimate of the impact on earnings, and how we use the treatment and comparison samples to compute nonexperimental impact estimates, such as the type of estimate computed in the WPRS evaluation.

In this evaluation, *the outcome of interest is the amount of earned income in the year after the quarter of random assignment*, which we refer to as "earnings" in the remainder of this report. The WPRS evaluation examined a variety of other outcome measures, including weeks of UI receipt. The estimated impacts on weeks of UI receipt reported in the WPRS evaluation are fairly consistent across states. The impact estimates range from zero weeks to a one-week reduction. However, the estimated impacts on earnings vary considerably across states, and this variation raises questions about the appropriateness of the regression model to the estimation of earnings impacts.

A. ESTIMATING IMPACTS WITH EXPERIMENTAL DATA

The key to this evaluation is the random assignment design of the JSA Demonstration. Because of random assignment, the average earnings of control group members provide an unbiased picture of what the earnings of treatment group members would have been had they not been assigned to services. This claim is supported by Table 6, which describes the baseline characteristics of all three samples. As one would expect under random assignment, treatment and control claimants have similar baseline characteristics. Table 1 shows that the mean profiling scores for treatment claimants and control claimants are 0.64 and 0.65, respectively. The means and frequencies for other baseline variables are also similar between the two groups.

The experimental impact estimate equals the average earnings for treatment group members minus the average earnings for control group members. The two samples are weighted according to Table 4 from Chapter II. Due to random assignment, the treatment-control difference in earnings provides an unbiased estimate of the impact of being assigned to the treatment.

B. ESTIMATING IMPACTS WITHOUT EXPERIMENTAL DATA

Unlike the JSA Demonstration, the WPRS evaluation lacked an experimental design and therefore lacked a randomly assigned control group against which the treatment group could be compared. In the WPRS evaluation, a comparison group of claimants who were not assigned to WPRS was selected to serve as a substitute for the control group. Because claimants were assigned to WPRS services based on their profiling scores, there are systematic baseline

TABLE 6: BASELINE VARIABLES FOR ALL THREE SAMPLES

Characteristics	Treatment Claimants	Control Claimants	Comparison Claimants	
			All	Profiling Score >= .40
Profiling Score (0 – 1)	0.64	0.65	0.45	0.48
Propensity Score (0 – 1)	0.86	0.86	0.09	0.13
Base-year earnings ($)	22,950	23,452	18,615	18,425
Age (years)	47	48	43	44
Sex:				
Male	52.2%	52.2%	55.7%	55.1%
Female	47.8%	47.8%	44.3%	44.9%
Race/Ethnicity:				
White	58.8%	57.9%	51.0%	47.2%
Black	13.4%	14.3%	15.7%	15.0%
Hispanic	26.5%	26.7%	32.0%	36.5%
Other	1.2%	1.1%	1.4%	1.3%
Education:				
No High School Degree	30.2%	33.3%	21.2%	28.0%
High School Degree	55.0%	49.0%	49.1%	50.5%
Associate Degree	7.1%	9.4%	13.7%	10.0%
Bachelor's Degree	7.4%	8.3%	13.1%	10.4%
Graduate School	0.3%	0.0%	2.4%	1.0%
Job Tenure:				
Less Than 1 Year	13.5%	10.7%	79.6%	73.1%
1 to 3 Years	22.3%	21.7%	12.0%	15.8%
3 to 10 Years	32.0%	32.1%	4.6%	5.9%
10 Years or More	32.3%	35.5%	3.8%	5.2%
Local Office:				
Clearwater	9.6%	10.4%	4.9%	4.2%
Davie	14.5%	17.6%	12.1%	10.7%
Ft. Lauderdale	15.9%	12.5%	25.6%	25.3%
Ft. Pierce	5.1%	4.7%	2.3%	1.6%
Hialeah	16.1%	12.0%	19.5%	22.5%
Lakeland	5.0%	4.8%	2.5%	1.9%
Orlando	14.7%	15.9%	13.6%	13.0%
Pensacola	4.5%	4.9%	2.5%	1.9%
St. Augustine	5.3%	4.2%	1.2%	0.8%
Miami	9.2%	13.0%	15.9%	18.1%
Sample Size	2,386	788	4,968	2,117

Note: The samples shown in Table 6 are weighted according to Tables 4 and 5.

differences between the treatment and comparison groups. In Connecticut, for example, the mean profiling score was 0.61 for treatment claimants and 0.49 for comparison claimants (Dickinson et al., 1999, Exhibit III-2).

In identifying the treatment and comparison samples for this evaluation, we mimic the rule by which claimants are assigned to WPRS. Therefore, like the treatment and comparison groups in the WPRS evaluation, the treatment and comparison groups in this evaluation are systematically different from each other. The baseline characteristics of both groups are displayed in Table 6. The average profiling score for treatment and comparison claimants are 0.64 and 0.45, respectively. Furthermore, the difference in "base-year earnings" between the two groups, $4,335, is quite large. Base-year earnings is a measure of earnings in four of the five quarters prior to applying for UI benefits. The treatment and comparison groups also differ on other baseline characteristics. Comparison claimants are more likely to be Hispanic, less likely to be white, and more likely to have a college degree than treatment claimants. The largest difference between the two groups is in the distribution of pre-displacement job tenure: four out of five comparison claimants had less than one year of job tenure, while fewer than one out of five treatment claimants had less than one year of tenure.

One difference between the treatment and comparison groups is based on the eligibility threshold for the demonstration. All treatment group members were eligible for the demonstration because they had profiling scores greater than 0.40, while only some comparison group members were eligible. Therefore, it is reasonable to ask whether dropping ineligible claimants would produce a better comparison group. The last column of Table 6 provides baseline characteristics for the comparison group members who were eligible for the demonstration. These results suggest that dropping ineligible claimants from the comparison group might produce a slightly better comparison group, but would *not* produce a group with similar baseline characteristics to the treatment group.

The large baseline differences between the treatment and comparison groups indicate that we should expect the unadjusted mean differences in earnings between the treatment and comparison groups to provide biased impact estimates. In this evaluation, we test two different approaches to addressing the large baseline differences: the regression methods used in the WPRS evaluation, and the matching methods designed for this evaluation.

C. REGRESSION METHODS USED IN THE WPRS EVALUATION

In the WPRS evaluation, impacts were estimated from the treatment and comparison samples using the following regression specification to adjust for baseline differences between the two groups:

(1) $EARNINGS = B_0 + B_1(TREATMENT\ INDICATOR) + B_2(X) + \varepsilon$

The treatment indicator equals one for treatment group members and zero for comparison group members. The control variables included in *X* varied across states in the WPRS evaluation due to data availability. However, the following baseline variables were included as control variables in all states: the profiling score; sex, race/ethnicity, age, and education; weekly UI benefit amount and weeks of UI entitlement; local office, unemployment rate, and quarter in which the claim was made; and pre-unemployment industry, occupation, earnings and job tenure. All of these variables were created from UI claims data.

In this evaluation, we estimate impacts from the treatment and comparison groups using the same regression model that was used in the WPRS evaluation to estimate impacts in most states. We estimate equation (1) and define the "regression-based impact estimate" as the least squares estimate of B_1.[7] Therefore, not only is the regression-based estimate generated from samples designed to mimic the WPRS samples, but it is also based on the most common regression specification from the WPRS evaluation.

D. MATCHING METHODS DESIGNED FOR THIS EVALUATION

For this evaluation, we designed alternative methods for estimating the impacts of WPRS. These methods are based on different ways of matching treatment claimants to comparable comparison claimants. Instead of adjusting for baseline differences between the treatment and comparison samples via regression analysis, we select subgroups of the comparison group called "matched comparison groups" that are designed to be more comparable to the treatment group than the entire comparison group. For each matched comparison group, impacts are estimated by the difference between the average earnings for treatment group members and the average earnings for matched comparison group members.

Each matched comparison group is selected via a matching rule to determine which comparison claimants could be matched to each treatment claimant. Therefore, a comparison claimant is selected for the matched comparison group if he or she can be "matched" to one or more treatment claimants with similar characteristics, where the definition of a match depends on the matching rule. The matching rules tested in this evaluation are described in the next section.

Rules for Matching Treatment Claimants to Comparison Claimants

The matching rules used to select matched comparison groups from the entire comparison group are defined by answers to the following two questions:

1. On which baseline variables must matching treatment and comparison claimants have similar values, i.e. which variables will be used in matching?

2. For matching claimants, how similar must the values of these variables be?

It would be desirable to define matched pairs of treatment and comparison claimants based on *all* of the baseline variables that influence earnings. However, if the number of such variables is large, there may be many treatment units without any matching comparison units. Therefore, we select a subset of the baseline characteristics to use in matching. Each matching rule contains four matching variables: sex, race/ethnicity, education, and one additional variable. The

[7] The treatment and comparison samples are weighted according to Tables 4 and 5 from Chapter II.

matching rules differ primarily in the choice of that additional variable. The three additional variables used to define the matching rules tested in this evaluation are as follows:

1. ***Profiling Score.*** Because claimants are assigned to WPRS based on their profiling scores, the most obvious way in which the average comparison claimant differs from the average treatment claimant is in the profiling score.[8] Therefore, we selected a matched comparison group with similar profiling scores to the treatment group.

2. ***Base-year Earnings.*** Because earnings is our outcome of interest, it would be desirable if the treatment and comparison groups had similar average earnings before applying for UI. Therefore, we selected a matched comparison group with similar base-year earnings to the treatment group.

3. ***Propensity Score.*** Treatment and comparison claimants differ in the probabilities (or propensities) that they would be assigned to services, and in the characteristics that determine these propensities. Therefore, we selected a matched comparison group with similar propensity scores to the treatment group. (We describe how these scores were computed later in the chapter.)

The differences between the four matching rules are based on the three variables listed above and on how those variables are used to define matches. These differences are summarized in Table 7. In three out of the four matching rules, we match each treatment claimant to the single comparison claimant who is "closest" to the treatment claimant in the key matching variable (among comparison claimants with the same sex, race/ethnicity, and education). To take advantage of the possibility that there may be many comparison claimants who are close matches for particular treatment claimants, the fourth matching rule allows each treatment claimant to match all comparison claimants who are sufficiently close to the treatment claimant on the key matching variable (among comparison claimants with the same sex, race/ethnicity, and education).

[8] The profiling score is a function of the variables used to predict the exhaustion of UI benefits, such as education, industry, occupation, and job tenure. Therefore, the comparison group should differ systematically from the treatment group in the distributions of these variables.

TABLE 7: DIFFERENCES BETWEEN MATCHING RULES

Matching Rule	Key Matching Variable	Matching Criterion Based on Key Variable
1	Profiling Score	Closest Match
2	Base-year Earnings	Closest Match
3	Propensity Score	Closest Match
4	Propensity Score	All Close Matches (tolerance of 0.1)

All four of our matching rules allow the same comparison claimant to match multiple treatment claimants. This flexibility is important when the comparison group is small, or when the treatment and comparison groups are so different that some subgroups have very few comparison units per treatment unit.

Estimating the Propensity Score

Of the three key matching variables, two are available in the data collected for the demonstration. Base-year earnings are computed to determine eligibility for UI and are therefore available from UI administrative records. Profiling scores were computed especially for the demonstration and are now computed to determine which claimants should be assigned to WPRS. However, in this evaluation, we computed an additional matching variable, the propensity score, from the available baseline variables for the following two reasons: (1) propensity scores are the most commonly used matching variables in the evaluation literature, and (2) the evidence described in Chapter I suggests that propensity score matching produced accurate estimates of the impacts of the National Supported Work Demonstration (Dehejia and Wahba, 1999).

The propensity score provides an estimate of the probability that a claimant would be assigned to WPRS. To compute propensity scores for treatment and comparison claimants, we first estimated a logit model of assignment to WPRS as a function of the three factors that affect assignment: the profiling score, the local office in which the claim was filed, and the time period when the claim was filed. Second, we used the estimated logit model to compute a propensity score for each treatment and comparison claimant.

The propensity scores reflect that the probability of being assigned to WPRS is larger for claimants with high profiling scores, claimants who applied for benefits in smaller offices (with relatively few other profiled claimants), and claimants who applied for benefits in months in which relatively few other claimants were profiled. Figure 1 shows the relationship between the profiling score and the estimated propensity score for selected local offices and months. As indicated in Chapter II, larger offices serve a smaller proportion of profiled claimants than smaller offices. Therefore, we chose one large office, Ft. Lauderdale, and one small office, St. Augustine, to illustrate that the probability of being assigned to WPRS services is higher in smaller offices.[9]

[9] The size difference between Ft. Lauderdale and St. Augustine is reflected in the average number of claimants profiled each week in each office (as shown in Table 1).

However, even within similar sized offices, the relationship between the profiling score and the propensity score varies. As shown previously in Table 1, Clearwater and Miami both profiled approximately the same number of claimants per week during the demonstration. However, for a given profiling score, the probability of being assigned to services was higher in Clearwater than Miami because the number of claimants with high profiling scores was larger in Miami.

Lastly, the probability that claimants would have been assigned to WPRS varied across the different weeks and months of the demonstration. For presentational purposes, Figure 1 shows the relationship between the profiling and propensity scores for two months only, July 1995 and September 1995. These two months were picked because assignment probability was highest in July and lowest in September, and this difference is reflected in Figure 1.[10]

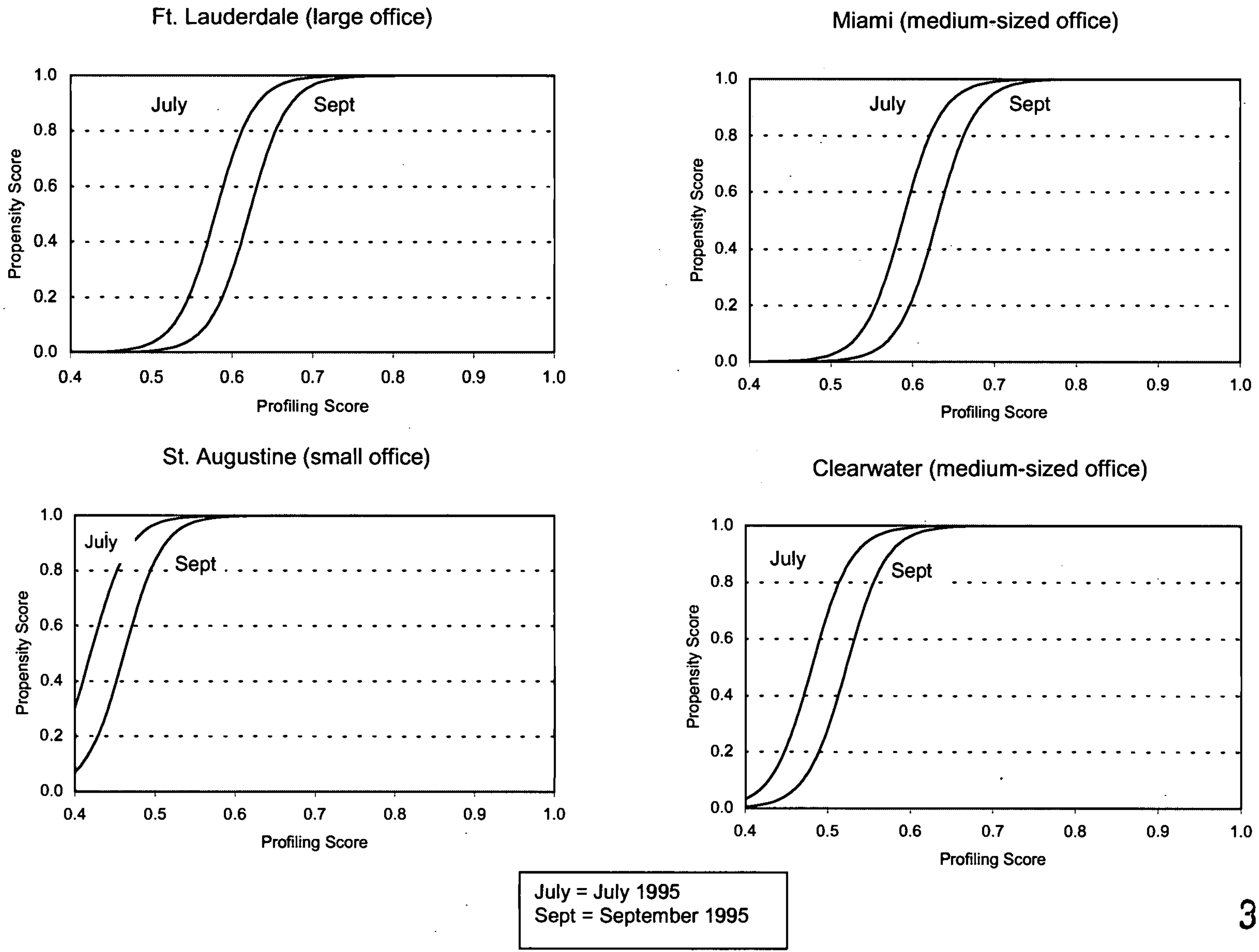
Success in Finding Acceptable Matches

To create matched comparison groups, it must be feasible to match treatment group members to comparison group members with similar characteristics. In settings where many of the treatment units cannot be matched, impact estimates based on the treatment units that can be matched may not be generalizable to the entire treatment group. Furthermore, if the inability to match many treatment units is an indication of large differences between treatment and comarison groups, then *no known method of estimating impacts from the two groups would be consistently accurate.*

Therefore, a matching rule is only useful in this evaluation if most treatment claimants can be matched to similar comparison claimants according to the rule. Whether most treatment claimants can be matched depends on the similarity between the two groups in the distribution of the matching variables. The potential for matching between the two groups on the basis of the profiling score is reflected in Figure 2. For each of the two samples, Figure 2 provides the number of claimants whose profiling scores fall within the following ranges: 0.4 – 0.5, 0.5 – 0.6, 0.6 – 0.7, and 0.7 – 1.0. This figure illustrates that the two distributions are very different from each other. For example, there are many more treatment claimants than comparison claimants with profiling scores greater than or equal to 0.7. However, there is considerable overlap in the two distributions: each of the four categories contains some claimants from each group. We address the difference in the two distributions by allowing each comparison claimant to match multiple treatment claimants (under all four matching rules).

[10] This difference suggests that the number of claimants with high profiling scores was lowest in July and highest in September.

FIGURE 1: PROPENSITY SCORE FUNCTIONS FOR SELECTED LOCAL OFFICES AND MONTHS



Ft. Lauderdale (large office)

Miami (medium-sized office)

St. Augustine (small office)

Clearwater (medium-sized office)

July = July 1995
Sept = September 1995

FIGURE 2: DISTRIBUTION OF PROFILING SCORES
(Treatment and Comparison Groups)

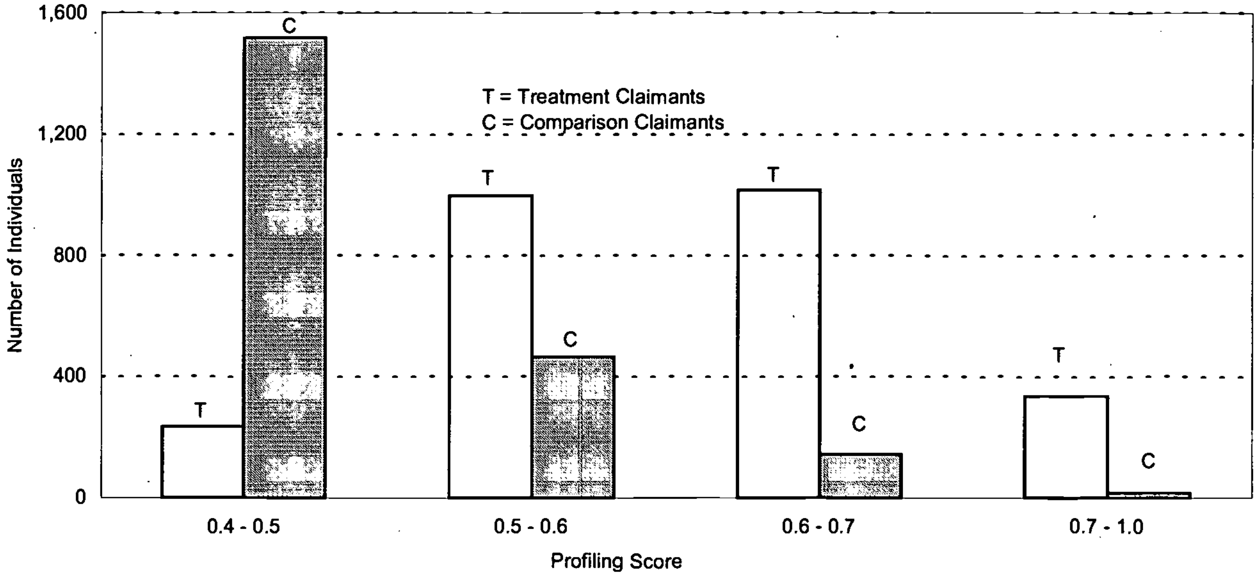
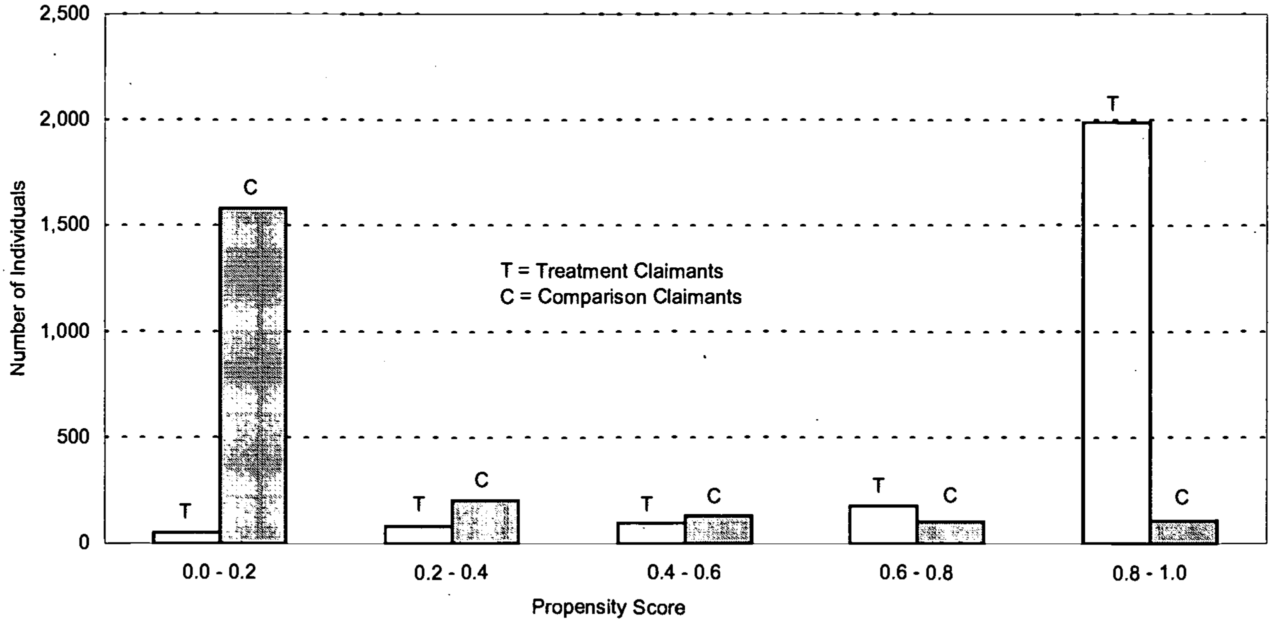


FIGURE 3: DISTRIBUTION OF PROPENSITY SCORES
(Treatment and Comparison Groups)



The overlap in the profiling score distributions between the treatment and comparison groups is due to variation in the WPRS assignment threshold across local offices and weeks of the demonstration. Within a particular local office and week, the assignment threshold perfectly distinguishes between treatment claimants, who have profiling scores above the threshold, and comparison claimants, who have profiling scores below the threshold. Therefore, it is impossible to match treatment claimants to comparison claimants with the same profiling score in the same office and week. However, because the profiling score threshold varies across local offices and weeks, it is possible to match treatment claimants to comparison claimants with the same profiling score *in different offices or weeks.*

Figure 3 provides the propensity score distribution separately for treatment and comparison claimants. Like the profiling score distribution, the propensity score distribution differs considerably between the two groups. Treatment claimants have higher average propensities than comparison claimants because they have higher average profiling scores. The difference in average propensity scores between the two groups is reflected in the tails of the distribution. For While comparison claimants vastly outnumber treatment claimants among those with propensity scores less than 0.2, treatment claimants vastly outnumber comparison claimants among those with propensity scores greater than 0.8.

Even with the large differences between the treatment and comparison groups in the distributions of the key matching variables, we were able to match most treatment claimants to at least one comparison claimant according to each of the four matching rules. The number of unmatched treatment claimants under each rule is provided in Table 8:

TABLE 8: UNMATCHED TREATMENT CLAIMANTS

Matching Rule	Tolerance Range for Matches	Percent of Treatment Units Unmatched
1	|Profiling score difference| ≤ 0.1	3.4%
2	|Base-year earnings difference| ≤ $1,000	4.4%
3 and 4	|Propensity score difference| ≤ 0.1	4.7%

The goal of matching is to create matched comparison groups that are comparable to the treatment group in ways that the entire comparison group is not. As shown earlier in Table 6 (and reproduced in the first two columns of Table 9), the entire comparison group has very different baseline characteristics from the treatment group. In Table 9, we provide evidence that in many ways, the matched comparison groups have similar baseline characteristics to the treatment group.

The first matched comparison group (based on matching rule 1) is selected to ensure that the profiling scores of matched comparison claimants are comparable to the profiling scores of treatment claimants. Table 9 shows the average profiling score is nearly identical for the two groups--0.64 for the treatment group and 0.63 for the first matched comparison group. Furthermore, the difference in the job tenure distribution between the two groups is much

TABLE 9: BASELINE VARIABLES FOR TREATMENT, COMPARISON, AND MATCHED COMPARISON GROUPS

Characteristics	Treatment Claimants	Comparison Claimants	Matched Comparison Claimants			
			Matching Rule 1	Matching Rule 2	Matching Rule 3	Matching Rule 4
Profiling Score (0 – 1)	0.64	0.45	0.63	0.43	0.59	0.59
Propensity Score (0 - 1)	0.86	0.09	0.70	0.11	0.84	0.83
Base-year earnings ($)	22,950	18,615	26,474	21,369	19,610	21,591
Age (years)	47	43	48	42	45	44
Sex:						
Male	52.2%	55.7%	52.5%	53.0%	54.9%	54.9%
Female	47.8%	44.3%	47.5%	47.0%	45.1%	45.1%
Race/Ethnicity:						
White	58.8%	51.0%	59.3%	59.9%	62.9%	62.9%
Black	13.4%	15.7%	13.3%	12.8%	13.8%	13.8%
Hispanic	26.5%	32.0%	26.8%	26.6%	23.3%	23.3%
Other	1.2%	1.4%	0.7%	0.7%	0.0%	0.0%
Education:						
No High School Degree	30.2%	21.2%	29.2%	28.6%	30.4%	30.4%
High School Degree	55.0%	49.1%	55.5%	56.7%	54.6%	54.6%
Associate Degree	7.1%	13.7%	8.7%	5.2%	11.5%	9.0%
Bachelor's Degree	7.4%	13.1%	6.6%	8.2%	3.5%	6.0%
Graduate School	0.3%	2.4%	0.0%	1.2%	0.0%	0.1%
Job Tenure:						
Less Than 1 Year	13.5%	79.6%	12.8%	80.5%	42.9%	32.1%
1 to 3 Years	22.3%	12.0%	32.6%	11.5%	24.3%	32.8%
3 to 10 Years	32.0%	4.6%	15.7%	5.4%	16.6%	15.1%
10 Years or More	32.3%	3.8%	38.9%	2.7%	16.2%	19.9%
Local Office:						
Clearwater	9.6%	4.9%	2.7%	5.8%	19.5%	13.1%
Davie	14.5%	12.1%	8.9%	13.6%	9.7%	10.3%
Ft. Lauderdale	15.9%	25.6%	33.2%	20.3%	3.1%	5.6%
Ft. Pierce	5.1%	2.3%	0.9%	6.6%	6.4%	7.9%
Hialeah	16.1%	19.5%	26.1%	16.8%	15.6%	17.0%
Lakeland	5.0%	2.5%	1.2%	7.1%	6.7%	7.5%
Orlando	14.7%	13.6%	6.7%	9.2%	9.0%	8.8%
Pensacola	4.5%	2.5%	3.0%	6.7%	10.7%	11.4%
St. Augustine	5.3%	1.2%	1.0%	3.0%	13.4%	11.9%
Miami	9.2%	15.9%	16.3%	10.9%	6.0%	6.5%

smaller than the difference between the treatment group and the entire comparison group. This result is not surprising because job tenure variables were included in the profiling model.

However, Table 9 reveals that some large differences remain between the treatment group and the first matched comparison group. Average base-year earnings differ considerably between the two groups. The local office distribution also differs considerably, but this is a predictable result of the evaluation's design. The first matched comparison group is disproportionately comprised of claimants from large offices because large offices contained a disproportionately large number of claimants who would not be assigned to WPRS despite having high profiling scores.[11]

The second matched comparison group (based on matching rule 2) is selected to ensure that the base-year earnings of matched comparison claimants are comparable to the base-year earnings of treatment claimants. Table 9 shows that while the difference in average base-year earnings between the treatment group and the entire comparison group is $4,335, the difference between the treatment group and the second matched comparison group is only $1,581. Therefore, matching rule 2 reduced the treatment-comparison difference in base-year earnings. It also almost eliminated the difference in the educational distribution between the two groups. However, matching rule 2 did nothing to reduce the large differences in average profiling scores and in the job tenure distribution.

The third and fourth matched comparison groups (based on matching rules 3 and 4) are selected to ensure that the propensity scores of matched comparison claimants are comparable to the propensity scores of treatment claimants. The difference in the average propensity score between the treatment group and the entire comparison group is 0.19 or 19 percentage points. The difference between the treatment group and either the third or the fourth matched comparison group is only 5 percentage points. Therefore, matching rules 3 and 4 reduce but do not eliminate the treatment-comparison differences in the average propensity score. Matching rules 3 and 4 also reduce but do not eliminate the treatment-comparison differences in the following distributions: race/ethnicity, education, and job tenure. Finally, unlike matching rule 1, matching rules 3 and 4 select a disproportionate share of comparison claimants from *small* offices for the following two reasons: (1) treatment claimants tend to have high propensities, and (2) comparison claimants in small offices tend to have relatively high propensities, and are therefore better candidates for matching than comparison claimants in large offices.

Estimating impacts based on the four matched comparison groups requires weights that account for variation in the sampling rate. As shown previously in Table 4, we assign weights to treatment claimants that reflect the sampling probabilities in the local offices where they applied for benefits. We assign weights to matched comparison group members that reflect the sampling probabilities *of the treatment group members to which they were matched.* The details of the weighting procedures are provided in Appendix A.

[11] As explained earlier, claimants who applied for benefits in a particular local office and week cannot be matched to comparison claimants with the same profiling score in the same office and week. Furthermore, the probability of being assigned to WPRS was lower in large offices than in small offices. Therefore, matching rule 1 selects a disproportionately large share of the matched comparison claimants from large offices.

CHAPTER IV: RESULTS

Chapter IV presents the results of the evaluation. This evaluation designed and implemented a test of different methods for estimating the impacts of WPRS programs based on nonexperimental samples like those used in the WPRS evaluation. This chapter shows the results of applying the methods described in Chapter III to the samples described in Chapter II. Based on data from the JSA Demonstration, the two main findings from the evaluation are given below:

1. The regression-based method used in the WPRS evaluation produced accurate impact estimates.
2. The matching methods tested in this evaluation produced less accurate impact estimates than the regression-based method.

Whatever general concerns can be raised about the reliability of regression methods to adjust for large differences between the treatment and comparison groups, the regression-based impact estimate we compute using the methods from the WPRS evaluation is very close to the experimental impact estimate. *Therefore, this evaluation provides no evidence that the regression methods used in the WPRS evaluation are unreliable.*

A. EXPERIMENTAL IMPACT ESTIMATES

The experimental impact estimate equals the difference in average earnings between the treatment and control groups. Recall that the treatment and control groups for this evaluation are subgroups of the demonstration's treatment and control groups who would have been assigned to WPRS had it been operating instead of the demonstration. The treatment-control difference in average earnings serves as a benchmark against which other estimates can be compared.

The experimental impact estimate is provided in column (1) of Table 10. As shown in panel A, the difference in average earnings between the treatment and control groups is $260. This estimate is almost exactly equal to the earnings impacts for the subgroup of demonstration participants in the top quartile of the profiling score distribution--a subgroup which may roughly approximate the claimants who would have been assigned to WPRS.[12] Decker et al. (2000) reported that the average earnings impacts were -$158, $804, and $139 for the three treatments in the demonstration. Since each treatment group has approximately the same number of individuals, the average impact across the three treatments is the simple average of the three impacts, or $262. Therefore, the experimental impact estimate of $260 is consistent with findings from the demonstration.

[12] This subgroup consists of the 2,257 treatment group members and the 754 control group members with the highest profiling scores. The experimental samples in this evaluation consist of 2,386 treatment group members and 788 control group members with the highest profiling scores *in the office and week in which they applied for benefits.*

Regression adjustments are often made in experimental evaluations to increase the precision of the impact estimates. Therefore, panel B shows the regression-adjusted difference in average earnings between the treatment and control groups. These adjustments are based on the same regression specification used to compute our regression-based estimates. The regression-adjusted experimental impact estimate equals $32 with a standard error of $583. Therefore, the regression adjustments slightly improve the precision of the impact estimates. Both the unadjusted and regression-adjusted experimental estimates fall within the range of the impact estimates reported in the WPRS evaluation.

TABLE 10: ESTIMATES OF IMPACTS ON EARNINGS
(Dollars)

Impact Estimate	Experimental (1)	Regression (2)	Matching 1 (3)	Matching 2 (4)	Matching 3 (5)	Matching 4 (6)
Unadjusted	260 (695)	-1,220 (396)	-3,440 (1,995)	-2,460 (523)	-111 (1,347)	-2,025 (1,233)
Regression-Adjusted	32 (583)	308 (695)	-1,516 (1,387)	1,139 (893)	-424 (1,242)	-1,162 (1,182)

NOTE: Impact estimates measure the average impact of assignment to treatment services for demonstration participants who would have been assigned to WPRS services. The unadjusted impact estimates are based on the raw treatment-control differences for column (1), and based on the raw treatment-comparison differences for columns (2) – (6). The regression-adjusted impact estimates are based on a linear model that includes the following control variables: the profiling score; sex, race/ethnicity, age, and education; weekly UI benefit amount and weeks of UI entitlement; local office, unemployment rate, and quarter in which the claim was made; and pre-unemployment industry, occupation, earnings and job tenure.

The usefulness of the experimental benchmark depends on how precisely it is measured. Table 10 shows that the estimated standard error on the experimental impact estimate is $695. This estimate is large enough that not only is the impact estimate insignificantly different from zero, but it is insignificantly different from any impact estimate that would seem credible given the evaluation of the demonstration and WPRS. Results from the demonstration indicate that impacts were measured more precisely in the demonstration than in this evaluation.[13] The difference can be attributed to two factors. First, the treatment-control differences in the demonstration were based on samples of approximately equal size, while the treatment-control differences in this evaluation are based on a treatment-control ratio of 3:1.[14] Second, the weights in this evaluation vary considerably across local offices. These weights are necessary to obtain unbiased estimates of the average impacts across the 10 offices, but the variation across local offices reduces the precision of the experimental benchmark estimate. Therefore, given the

[13] For one of the three treatments (Individualized Job Search Assistance Plus Training), the estimated impact for females was $799. The estimated standard error of this impact estimate was $407, which is smaller than the estimated standard error of the experimental benchmark estimate in this evaluation. The smaller standard error is not due to a larger sample, because the sample of females used to compute the impact estimate of $799 was larger than the sample of treatment and control claimants used to compute the experimental benchmark estimate.

[14] We combined the three treatment groups for this evaluation because separate impact estimates for each treatment would have been considerably less precise.

relative imprecision of the experimental benchmark estimate, it is challenging to detect differences between the experimental benchmark and any of the nonexperimental estimates.

B. REGRESSION-BASED IMPACT ESTIMATES

The regression-based impact estimate is based on samples and methods designed to mimic the samples and methods from the WPRS evaluation. As shown in column (2) of Table 10, panel B, the regression-based impact estimate equals $308 with an estimated standard error of $695. This impact estimate is very close to the experimental benchmark of $260. Given the relative imprecision of both estimates, we cannot conclude with confidence that the experimental and regression-based methods would consistently produce similar impact estimates. However, given the small difference between the two estimates, there is very little opportunity for the impact estimates based on matched comparison groups to improve upon the regression-based estimates in this evaluation.

It is worth noting the contribution of the regression adjustments to the performance of the regression-based estimate. The unadjusted treatment-comparison difference in earnings is -$1,220, as shown in column (2) of Table 10, panel A. Therefore, the unadjusted difference is very different from the experimental benchmark estimate, and the regression adjustments are responsible for the small difference between the experimental benchmark and the regression-based estimate.

C. IMPACT ESTIMATES BASED ON MATCHED COMPARISON GROUPS

The estimates of earnings impacts presented in this section are based on the treatment group and the four matched comparison groups, which were described in Chapter III. For each matched comparison group, Table 10 provides two impact estimates: the unadjusted difference and the regression-adjusted difference in average earnings between the treatment group and the matched comparison group. Based on these results, we conclude that most of the matching methods do not perform very well when applied to data from the demonstration, and that none of the matching methods outperform the regression method.

First, we focus on the unadjusted impact estimates to isolate the effects of matching. For three of the four matching rules, the unadjusted earnings impact is *further* from the experimental benchmark than the unadjusted earnings difference between the treatment group and the entire comparison group. *Therefore, for most of the matching rules, the matched comparison group behaves even less like a randomly selected control group than the entire comparison group.* This finding is surprising given that as shown previously in Table 9, all four of the matched comparison groups seem more similar to the treatment group than the entire comparison group. Only for matching rule 3--which matches treatment claimants to the comparison claimant with the closest propensity score--is the unadjusted impact estimate closer to the experimental benchmark than the unadjusted earnings difference between the treatment and comparison groups. However, the unadjusted earnings difference between the treatment and the third matched comparison group is still further from the experimental benchmark than the regression-based estimate. Therefore, the regression method seems to outperform even the best performing matching method.

Like the experimental and regression-based impact estimates, the impact estimates based on the four matched comparison groups are shown in Table 10. The first matching rule selects the comparison claimant with the closest profiling score to each treatment claimant. Based on the first matched comparison group, the estimated impact on earnings is -$3,440. This impact is large, negative, and far from the experimental benchmark of $260. Furthermore, the estimated standard error equals $1,995, which is almost three times as large as the estimated standard error of the regression-based estimator. These results suggest that the estimates based on the first matched comparison groups are more biased and more imprecise than the regression-based estimate. However, because of the large standard errors on the impact estimates, the differences between the two impact estimates and between each impact estimate and the experimental benchmark are not statistically significant.

The poor performance of the first matched comparison group can probably be attributed to the disproportionately large number of matched comparison group members from Hialeah and Ft. Lauderdale--the two largest local offices in the demonstration. For reasons described in Chapter III, large local offices contain a disproportionate share of comparison claimants with high profiling scores, and who would therefore be good matches for treatment claimants. Table 9 shows that Hialeah and Ft. Lauderdale are over represented in the first matched comparison group relative to the treatment group. Furthermore, probably because earnings levels tend to be relatively high in the areas served by these two offices, Table 9 shows that the average base-year earnings are considerably higher for the first matched comparison group than for the treatment group (difference of $3,524). Therefore, it is not surprising that in a year after claimants applied for benefits, average earnings are considerably higher for the first matched comparison group than for the treatment group (difference of $3,440). The poor performance of the first matched comparison group can be attributed to a high prevalence of matching treatment claimants to comparison claimants from areas of the state where earnings are relatively high.

Therefore, the results from the first matched comparison group suggest that selecting matched comparison group members with similar demographic characteristics and similar pre-unemployment job characteristics is not sufficient to create a good comparison group. If treatment and matched comparison claimants live in areas with different average earnings, the earnings of matched comparison claimants will be systematically different from the earnings that treatment claimants would receive if they had not been assigned to reemployment services. These results are consistent with findings in Heckman et al. (1997), but are even more striking because all treatment and comparison claimants reside in the same state: the spatial mismatch between the two groups is limited to different areas within the state of Florida. Therefore, when the average level of the outcome varies across local areas, it may be very important to match treatment units to comparison units that reside in the same area.

The second matching rule selects the comparison claimant with the closest base-year earnings to each treatment claimant. Based on the second matched comparison group, the estimated impact on earnings is -$2,460. Like the impact estimate based on the first matched comparison group, this estimate is far from the experimental benchmark. However, the explanation for poor performance is probably somewhat different for the second matching rule than for the first rule. Relative to the first matched comparison group, the second matched comparison group is more similar to the treatment group in average base-year earnings and the distribution of claimants across local offices, as shown in Table 9. However, Table 9 also shows that the second matching rule fails to reduce the difference between treatment and comparison groups in the average profiling score, average age, and the job tenure distributions. Claimants in the second matched comparison group tend to be considerably younger, have lower profiling scores, and have fewer years of job tenure prior to unemployment than treatment claimants. These differences indicate that claimants in the second matched comparison group lack many of the characteristics of displaced workers--the individuals that the demonstration and WPRS were designed to serve. Therefore, the second matched comparison group is not a good comparison group for this evaluation, and we should not be surprised by its inability to generate an impact estimate close to the experimental benchmark.

The third matching rule selects the comparison claimant with the closest propensity score to each treatment claimant. Based on the third matched comparison group, the estimated impact on earnings is -$111. This estimate is much closer to the experimental benchmark than the estimates based on the first and second matched comparison groups. However, it is still further from the experimental benchmark than the regression-based estimate.

The fourth matching rule selects the comparison claimants with similar propensity scores to each treatment claimant, allowing each treatment claimant to match multiple comparison claimants. Based on the fourth matched comparison group, the estimated impact on earnings is -$2,025. The large difference between this estimate and the estimate based on the third matched comparison group is surprising for two reasons: the two matching rules are very similar, and the two matched comparison groups have similar characteristics, as shown in Table 9. Therefore, while the evidence in Table 9 is useful in predicting the poor performance of the first two matched comparison groups, it is not useful in predicting the poor performance of the fourth matched comparison group relative to the third. However, the impact estimates based on both the third and fourth matched comparison groups are measured imprecisely, and the difference between the two impact estimates is insignificantly different from zero.

Thus far, the analysis of matching methods has focused on the unadjusted impact estimates. However, Table 10 also shows the regression-adjusted impact estimates using each of the four matched comparison groups. The regression-adjusted impact estimate is larger (less negative or more positive) than the unadjusted estimate for three out of four matched comparison groups, and the regression adjustments move the impact estimates closer to the experimental benchmark for each of these three groups. However, for the third matched comparison group, the regression-adjusted impact estimate is smaller than the unadjusted estimate, and the regression adjustments move the impact estimate further from the experimental benchmark. Therefore, the regression-adjustments do not always generate an impact estimate that is closer to the experimental benchmark than the unadjusted estimate. Furthermore, the combination of propensity score matching (matching rule 3)--the best performing matching rule--with regression adjustments does not perform better than either propensity score matching or regression adjustments alone.

All four matching rules selected matched comparison groups with different local office distributions than the treatment group. As described earlier, this difference may be responsible for the poor performance of the first matching rule (and partially responsible for the poor performance of other matching rules). In an attempt to eliminate this difference, we tested a modification to our matching rules that required that treatment claimants be matched to comparison claimants who applied for benefits in the same local office. However, as described earlier, there is little opportunity for matching within local office under matching rule 1 because of the rule by which claimants are assigned to WPRS. Furthermore, this modification to our matching rules leaves many treatment claimants unmatched. The percent of treatment claimants that cannot be matched to comparison claimants in the same local office is 28 percent for matching rule 1, 31 percent for matching rule 2, and 40 percent for matching rules 3 and 4. Therefore, we cannot impose this additional requirement without leaving a large proportion of treatment claimants unmatched.

D. INTERPRETATION OF RESULTS

The results of this evaluation indicate that when the regression methods used in the evaluation of WPRS are applied to samples from the Job Search Assistance Demonstration, the resulting estimate of the impact on earnings is very close to the experimental benchmark estimate. The imprecision of the impact estimates raises the question of whether the regression methods would perform equally well in other similar samples. However, this evaluation provides no evidence that the regression method used to estimate earnings impacts in the WPRS evaluation is unreliable.

Most of the matching methods proposed as alternatives to regression methods performed poorly: the resulting impact estimates were typically far from the experimental benchmark estimate. There are three general problems that can lead matching methods to perform poorly:

1. Many treatment group members cannot be matched.
2. Matched comparison group members are different from treatment group members in observed characteristics that are not included in the matching rule.
3. The treatment and matched comparison groups have different unobserved characteristics that are related to the outcome measure and therefore bias the impact estimates.

In this evaluation, the poor performance of the matching methods is primarily due to problem 2. Problem 1 is not a severe problem in this evaluation: Table 8 showed that most treatment claimants matched at least one comparison claimant. Problem 3, which is the focus of much attention in the evaluation literature, is not applicable to this evaluation because the process by which individuals are assigned to services is based exclusively on observed variables. Actual participation in services may be influenced by the unobserved characteristics of claimants, but this evaluation measures the impact of being assigned to services: it does *not* attempt to measure the impact of participation.

Table 9 illustrates that problem 2 is a severe problem in this evaluation. Each matching rule generates a matched comparison group that is very different from the treatment group in the distribution of claimants across local offices. The matched comparison groups are also different from the treatment group in the distributions of other variables. The most obvious solution is to eliminate observed differences through modifications to the matching rule. However, as described earlier, requiring matched pairs of claimants to have applied for benefits in the same local office leads to problem 1: a large minority of treatment claimants cannot be matched to comparison claimants in the same local office. Therefore, it is at least difficult and perhaps impossible to avoid all three potential matching problems in this evaluation.

These results should not be interpreted as evidence that matching does not work *in general*. The impact estimates in this evaluation are not measured very precisely for reasons discussed in section A. Therefore, the differences in impact estimates on which we must infer the relative performance of different methods are typically not significantly different from zero. Also, it may be possible to avoid all three potential matching problems in some evaluation settings. Therefore, the most appropriate lesson to be learned from the poor performance of the matching methods we tested is that selecting a credible comparison group based on a small number of matching variables can be very challenging.

The good performance of the regression method used in the WPRS evaluation raises the following question: could its performance have been measured without the aid of a randomized control group like the one from this evaluation? To a large extent, the poor performance of the four matching methods could be predicted based on comparisons between the matched comparison groups and the treatment group. The matching rules we tested were effective at reducing the differences between the treatment and comparison groups in some *but not all* of the baseline variables used to describe the samples, and this assessment does not require a randomized control group. It is unclear if similar assessments can be made for regression methods. Specification tests for regression models exist, but it is unclear whether these tests answer the question of whether a regression model applied to nonexperimental data will produce accurate impact estimates.