Do Regression-Adjusted Performance Measures for Workforce Development Programs Track Longer-Term Program Impacts? A Case Study for Job Corps

Peter Z. Schochet, Mathematica Policy Research
Jane Fortson, Mathematica Policy Research

October 2012

Draft

Abstract

Since the 1970s, major workforce development programs administered by the U.S. Department of Labor have used performance management systems to measure program performance. Yet, there is scant evidence about the extent to which these performance measures are correlated with longer-term program impact (value-added) estimates. This article examines this issue using extensive data from a large-scale experimental evaluation of Job Corps, the nation's largest training program for disadvantaged youths. At the time of the evaluation, most Job Corps performance measures did not account for differences in the characteristics of students served; since that time, performance measures have increasingly been adjusted for observable differences in student characteristics. Thus, we used detailed baseline survey, program intake, and local area data to regression adjust center performance measures and compared them to center-level impacts from the experiment. We find that while regression adjustment changed somewhat the center performance measures, they are not correlated with the impact estimates. Thus, there remain important unobserved differences between students attending different centers that are correlated with key participant outcomes.

INTRODUCTION

Since the 1970s, major federal workforce development programs administered by the U.S. Department of Labor (USDOL) have used performance management systems to measure program performance at the national, state, and local levels. The Comprehensive Employment and Training Act (CETA) Amendments of 1978, the Job Training Partnership Act (JTPA) of 1982, and the Workforce Investment Act (WIA) of 1998 each specified that program outcomes for USDOL workforce programs be measured and tracked. Many of these initiatives started prior to the Government Performance and Results Act of 1993 that required all federal agencies to establish performance goals and to measure performance against these goals. More recently, the Office of Management and Budget mandated in 2002 that all federal agencies use the Program Assessment Rating Tool to measure program performance.

The performance measures used for USDOL workforce programs have typically included program inputs (such as program enrollment), processes (such as the level and quality of program implementation), and, more often, short-term outputs (such as employment rates and earnings soon after program exit) and pre-post changes in outputs (Blalock & Barnow, 2001; Courty et al., 2011). Survey-based customer satisfaction measures and cost-per-client measures have also been used to measure performance. In 2005, under WIA, USDOL implemented common performance measures for adult, dislocated worker, and youth workforce programs with similar goals.

Performance measures are useful for monitoring compliance with program rules and for comparing short-term participant outcomes to pre-set performance standards. They can help hold program managers accountable for program outcomes and can also help foster continuous program improvement. However, a performance management system does *not* necessarily provide information on the *causal* effects of a program in improving participants' outcomes relative to what

they would have been in the absence of the program. Assessing causality requires estimates of program *impacts* based on rigorous impact evaluations. Due to the cost and time of obtaining impact estimates, however, it is not typically feasible to use this value-added approach as part of an ongoing performance measurement system. Consequently, an important empirical question is the association between performance measures and program impacts—especially in an age of scarcity, where available funds for ongoing impact evaluations are likely to be limited.

One approach for potentially improving the association between performance measures and program impacts is to adjust the performance measures for differences across performance units in participant characteristics and local economic conditions. To the extent that these confounding factors are correlated with performance measures (such as labor market outcomes), adjusting the performance measures using available data could more accurately reflect the success of program practices that are under the control of program managers. In essence, regression adjustment is a nonexperimental approach that can help "level the playing field." Regression adjustment can also reduce "cream-skimming" behavior where programs focus on providing services to participants who are most likely to succeed in the labor market, thereby improving program performance but not necessarily program impacts (Barnow, 2009; Heckman & Smith, 2011; Heinrich, 2004).

Thus, we would expect the relationship between program performance measures and experimental program impact estimates to be strongest for adjusted performance measures, making this the fairest test for assessing the impact-performance association. Barnow (2000) and Heckman, Heinrich, and Smith (2002) analyzed the relationship between adjusted performance measures and impacts using experimental data from the 16-site National JTPA Study (Orr et al., 1996). Both studies found no relationship between local programs' adjusted performance measures and longer-term impacts on employment and earnings. This literature is small, however, because only a limited

number of experimental evaluations of major USDOL workforce programs have been conducted, and JTPA programs in the 1980s and 1990s were the only large-scale USDOL workforce programs that ever used fully-adjusted performance standards (Barnow & Heinrich, 2010; Siedlecki & King, 2005).[1] Currently, under WIA, performance standards for state and local areas are negotiated between USDOL and state and local officials, although USDOL is currently considering options for using regression-adjusted performance measures (for example, using variants of Michigan's Value-added Performance Improvement System described in Barnow, 2006). As described below, the Job Corps performance measurement system now incorporates several model-based standards, providing some adjustment.

This article adds to the literature by examining the extent to which adjusted performance measures yield accurate assessments of program impacts using data from the National Job Corps Study (NJCS), a large-scale experimental evaluation of Job Corps, the nation's largest, most comprehensive federal education and job training program for disadvantaged youths. Job Corps, which is administered by USDOL, serves young men and women between the ages of 16 and 24 in Job Corps centers, where participants typically reside for an average of 8 months and receive intensive vocational training, academic education, and other services.

Job Corps is an important case study for such an analysis for several reasons. First, Job Corps has used a rigorous performance standards system since the late 1970s that has been a prototype for other government programs. Second, Job Corps is a performance-driven program; contract renewals to operate Job Corps centers as well as financial rewards to program staff are tied to measured

[1] There is a larger literature that generally shows a weak relationship between performance measures and impacts in welfare-to-work programs (Bartik 1995, Heckman, Heinrich, & Smith, 2002, and Hill, 2003 provide reviews).

center performance. Thus, program performance matters; key programmatic decisions are made based on how well centers perform along these measures.

Our study builds on the literature in several important ways. First, we compare findings for three sets of performance measures—unadjusted (which were used by Job Corps at the time of the NJCS), minimally adjusted (using program intake data), and more fully adjusted (using NJCS baseline survey data). With this approach, we are able to not only assess the relationship between performance measures and impacts but also investigate the degree to which adjustment changes that relationship. Second, the NJCS was the first experimental evaluation of a federal employment and training program in which all program applicants nationwide were subject to random assignment during the sample intake period. Thus, our results generalize to the full Job Corps program, whereas the results from previous studies using the JTPA data pertain only to a small number of purposively selected study sites. Third, our analysis is based on data covering a longer time period than previous studies—48 months after random assignment compared to 30 months for the National JTPA Study. Finally, we use a new approach to estimate center-level impacts on key outcomes using accurate predictions of program intake staff, made at the time of program application, about the centers to which treatments and controls would likely be assigned. This approach allows us to estimate representative impacts for each of 100 Job Corps centers nationwide that were in operation during the NJCS, and hence, allows us to look at the association between center-level impact estimates and center-level performance measures.

The main impact analysis for the NJCS examined average treatment-control differences on key outcomes and found that Job Corps improved education and training outcomes (such as the receipt of General Educational Development [GED] and vocational certificates and time spent in school), significantly reduced criminal activity, and improved earnings and employment outcomes in the two

years after program exit, although the longer-term analysis did not demonstrate that impacts were sustained beyond the two-year period (Schochet, Burghardt, & McConnell, 2008). Using NJCS data, Schochet and Burghardt (2008) also found that center-level impacts on key outcomes were not associated with the aggregate *unadjusted* center performance measure used by Job Corps. Students in higher-performing centers had better outcomes; however, the same pattern was observed for the control group members who would have been assigned to those centers.

This article extends the work of Schochet and Burghardt (2008) to examine whether *adjusting* performance measures for student and local area characteristics results in positive statistical associations between center-level performance measures and center-level impacts. We also examine the performance-impact relationship for each performance measure component, not only the aggregate performance measure that was used by Schochet and Burghardt (2008). We address the following research questions:

- Are Job Corps center performance rankings changed by regression adjustment?
- To what extent are regression-adjusted performance measures better able to distinguish between centers with larger impacts and those with smaller impacts?
- Are there specific performance measures that are more associated with impacts than others or the aggregate (overall rating) measure?

Our main finding is that although regression adjustment changes somewhat the performance rankings of centers, the adjusted performance ratings remain *uncorrelated* with center-level impacts. These results are consistent with the earlier literature cited above.

The remainder of this paper is in ten sections. In the next section, we provide an overview of Job Corps, and then provide an overview of the NJCS. In the following section, we discuss the Job Corps performance measurement system as it operated at the time of the study, and then discuss our data sources and analytic approach for examining the relationship between measured center

performance and impacts. The ensuing three sections present our empirical findings, followed by a discussion of possible reasons for our findings. Finally, we present our conclusions.

OVERVIEW OF JOB CORPS

Job Corps is the nation's largest, most comprehensive education and job training program for disadvantaged youths. It serves youths between the ages of 16 and 24, primarily in a residential setting. The program's goal is to help youths become more responsible, employable, and productive citizens. Each year, it serves more than 60,000 new participants at a cost of about $1.5 billion, which is more than 60 percent of all funds spent by USDOL on youth training and employment services. USDOL administers Job Corps through a national office and six regional offices.

Job Corps services are delivered in three stages: outreach and admissions (OA), center operations, and placement. OA counselors recruit students, inform them about the program, and ensure that they meet eligibility criteria. Center operations, which are the heart of the program, involve vocational training, academic education, residential living, health care, and a wide range of other services, including counseling, social skills training, health education, and recreation. At the time of the NJCS, these comprehensive services were delivered at 110 Job Corps centers nationwide. Most centers are operated by private contractors, although about one-quarter are operated by the U.S. Departments of Agriculture and of the Interior. After the youths leave the program, placement agencies help participants find jobs or pursue additional training.

Most Job Corps students reside at the Job Corps center while training, although about 12 percent are nonresidential students who reside at home. Enrollment in Job Corps does not have a fixed duration (duration is eight months on average, but varies widely). The program has a distinctive open-entry, open-exit educational philosophy, in which instruction is individualized and self-paced. At the time of the NJCS, Job Corps offered vocational training in more than 75 trades,

and a typical center offered training in 10 or 11 trades. Job Corps's academic education aims to alleviate deficits in reading, math, and writing skills and to provide a GED certificate. Job Corps has a uniform, computer-based curriculum for major academic courses.

OVERVIEW OF THE NJCS

USDOL sponsored the NJCS in 1993 to examine the effectiveness of the Job Corps program. The evaluation measured the program's average impacts on participants' employment and related outcomes, and assessed whether the value of the program's benefits exceeds its costs (Schochet, Burghardt, & McConnell, 2010). The NJCS was a national experimental evaluation where from late 1994 to early 1996, nearly 81,000 young people nationwide were randomly assigned to either a treatment group, who were allowed to enroll in Job Corps, or a control group, who were not allowed to enroll for a period of three years. NJCS findings are based on the comparisons of the outcomes of 9,409 treatments in the research sample and 5,977 controls using survey data collected during the four years after random assignment and administrative earnings data maintained by the Social Security Administration (SSA) covering nine years. The remaining eligible applicants (about 65,600) during the study period were randomly assigned to the treatment group that was not part of the research sample.

THE JOB CORPS PERFORMANCE MEASUREMENT SYSTEM

The Job Corps performance measurement system gathers data that are used to rate Job Corps centers on the outcomes of their participants. At the time of NJCS—Program Years (PYs) 1994 to 1996—the Job Corps performance measurement system included eight or nine measures in three areas: (1) *program achievement* measures, including reading gains, math gains, the rate of attainment of a GED certificate, and the vocational completion rate; (2) *placement* measures, including the placement

rate, the average wage at placement, the percentage of full-time placements, and the percentage of quality placements (defined as the percentage of placements in jobs that matched the area of training); and (3) *quality/compliance* measures, including one ARPA measure developed from observations made by regional office monitors during center reviews.

At the time of the NJCS, Job Corps minimally adjusted centers' performance standards for the GED completion rate and average wage rate measures, but not for other measures. The GED model accounted for state differences in GED requirements, and the wage model controlled for differences in prevailing wages across geographic areas. Other measures were *not adjusted* for the characteristics of students or their local areas.

For each measure, a center's score was compared to a fixed standard (or adjusted standard for the GED and wage rate measures), and the performance measure was the percentage of the goal that was met. These percentages were then weighted to yield a summary measure of how well a center performed relative to its standards. Table 1 summarizes the different performance measures and indicates the program years in which each component is available. Schochet and Burghardt (2008) provide more details on the measures, the samples used to calculate each component, and the weights used to calculate the summary performance measure.

Today's Job Corps performance measurement system is similar to the system in place at the time of the NJCS. Thus, our results are likely to be relevant to the system today. However, a handful of details are different. In particular, the current system rates centers along 14 different dimensions, rather than 8 or 9, including additional measures of postplacement outcomes (such as employment and earnings at 12 months after placement). In addition, the current system uses model-based standards—rather than national standards—for a larger share of performance measures. For PY 2010, several different performance measures used model-based goals: the high school

diploma/GED attainment rate, the combined high school diploma/GED/career technical training attainment rate, the average literacy (numeracy) gain, the graduate average wage at placement, and the graduate six-month average weekly earnings. Because the current system relies more on model-based standards, it is arguably closer to a system with adjusted standards.

DATA SOURCES

Our analysis draws on data from seven sources: (1) Job Corps performance measurement system data, (2) NJCS baseline survey data, (3) baseline data from program intake forms, (4) data on local area characteristics, (5) Job Corps center characteristics, (6) Job Corps intake counselors' predictions of center assignment for NJCS participants, and (7) follow-up data from the NJCS.

Performance Measurement System Data

Our study used performance data covering program years (PYs) 1994 to 1996, because this was the period when NJCS treatment group members were enrolled in Job Corps centers. Because the treatment group participated in Job Corps over a relatively long period, the main performance measures for our analysis were constructed using three-year (and two-year, where applicable) averages of the performance ratio (i.e., the center's score relative to its standard) for the overall rating and for each component. These average performance ratios provide a measure of average (typical) center performance to which participants were exposed. In addition, we used data on each performance component from each year to examine the robustness of study findings. We restricted our analysis to the 100 Job Corps centers with performance measure data in all three years and sufficient sample to estimate center-level impacts (as described below).

Baseline Data From the NJCS and Program Intake Forms

To regression-adjust the center-level performance measures, we used data on the baseline characteristics of NJCS sample members. These data came from two sources: the ETA-652 program application form and the NJCS baseline survey. The program intake data contain baseline data collected on all program participants at intake to help determine an applicant's eligibility for Job Corps. The NJCS baseline survey, which had a 95 percent response rate, was conducted soon after random assignment and is much more detailed and complete than the program intake data (see Tables 2 and 3). Importantly, the program intake data items are similar to the data items that USDOL used in the 1980s and 1990s to adjust performance standards for JTPA youth programs.[2] We restricted our analysis to the 14,653 participants with full NJCS baseline data.

By using both sets of characteristics, we were able to see what regression adjustment achieves under two different scenarios. Using the program intake data, we were able to assess the influence of regression adjustment using data that would be available to program staff (the likely scenario). Using the NJCS baseline data, we were able to see what regression adjustment could achieve if a wide range of characteristics were measured, and measured well, at baseline (the best-case scenario).

Local Area Characteristics

Because local area characteristics may influence outcomes and are not under the control of program managers, we also accounted for differences in local area characteristics across students served by different Job Corps centers. To identify these characteristics for each NJCS participant, we matched pre-program participant zip codes from the program intake forms to area characteristics

[2] The data items used to adjust the JTPA youth employment rate performance standard included gender, age, high school graduation status, race, whether received cash welfare, whether received SSI, whether lacks significant work history, offender status, current labor force status, whether unemployed 15 or more weeks, the local unemployment rate, and the percent of families in the local area with incomes below the poverty level.

from the 2008 Area Resource File (ARF), a compilation of data from numerous data sources (U.S. Department of Health and Human Services, 2010). This study used the ARF local area characteristics listed in Table 4. Of the 14,653 NJCS participants with baseline data, we were able to identify local area characteristics for 14,542 participants.

Job Corps Center Characteristics

In our sensitivity analysis, we constructed regression-adjusted performance measures that controlled for key center characteristics: size (number of Job Corps participants), operator (private or Federal agency), and geographic region (of 10 possible regions). This approach adjusts for systematic differences in performance across centers of different types, and holds center managers harmless for these differences.

Intake Counselors' Predictions of Center Assignment

Our analysis required data on predictions by intake counselors regarding likely center assignments of NJCS participants. We used these predictions to match NJCS study participants to Job Corps centers. These predictions were required because nearly all control group members and about 27 percent of treatment group members did not attend Job Corps centers. Thus, without counselors' predictions of the center of assignment, we would not be able to identify treatment non-participants and control group members for each center. The predictions were collected at intake (which was prior to random assignment) and are thus available for both treatments and controls. They were found to be about 95 percent accurate (as determined by comparing predicted to actual center assignments for treatment group enrollees) and are available for 93 percent of the sample.

Follow-up Data From the NJCS

Center-level impacts were calculated using outcome data from the NJCS 12-, 30-, and 48-month follow-up interviews. Table 5 shows the education, arrest, and earnings outcome measures that were used in our analysis; these were the key outcomes used for the NJCS impact study. The earnings outcomes pertain to calendar years 1997 and 1998, roughly three and four years after random assignment, respectively. All outcome measures pertain to the full sample, except the GED attainment rate, which pertains to the 80 percent of the sample who did not have a high school credential at baseline.[3]

These outcome measures conceptually align with many of the Job Corps performance measures discussed above. However, there are important differences between the performance and outcome measures that may weaken their association. These include (1) the mode of data collection; (2) the pools used to construct the measures (for example, the Job Corps vocational completion rate is measured using only those who remained in a Job Corps center for at least 60 days, whereas the NJCS vocational completion rate pertains to the full sample); and (3) the fact that Job Corps performance measures include all center enrollees, compared to a random sample of enrollees for the NJCS outcome measures. These issues are discussed in more detail below.

The follow-up analysis sample includes NJCS participants who completed the 48-month interview, a total of 11,313 NJCS participants (6,288 treatments and 4,485 controls). About 81 percent of the treatment sample and 78 percent of the control sample responded to the 48-month interview. Of the follow-up participants, 10,409 were predicted to attend the 102 centers in our performance measure data. However, in two centers, there are ten or fewer NJCS participants in the

[3] The NJCS also collected SSA earnings records, but SSA did not release individual-level data for study sample members, instead returning output from provided computer programs that they ran to estimate impacts. The data agreements with SSA have long expired. Thus, these data could not be used for the present analysis.

follow-up sample. Impacts estimated using such small samples are likely to be very noisy estimates of the true program impact; because of this concern, we focused our analysis on the 100 centers (and the associated NJCS participants) with more than ten predicted participants in the follow-up sample. Our results are robust to the inclusion of the two very small centers (not shown).[4]

ANALYSIS METHODS

This section describes our methods for developing adjusted performance measures and center-level impacts. We link the discussion to the steps that are summarized in Figure 1.

Developing Adjusted Performance Measures

After gathering data on performance measures [**Step 1 of Figure 1**], we linked data on the characteristics of participants to Job Corps centers using predicted center assignments. We then calculated center-level averages of participant characteristics using the NJCS baseline survey data, program intake data, and ARF data on local area characteristics [**Step 2 of Figure 1**].

When calculating center-level average characteristics, we transformed categorical variables into indicators or groups of indicators. For each characteristic, we calculated the center average among predicted center participants with nonmissing data for that characteristic. For items with large numbers of missing values, we constructed center-level variables signifying the proportion of sample members with missing values. Center-level averages were weighted by the NJCS baseline weight, which accounts for differences in sampling and survey response probabilities (see Schochet, 2001). The analysis sample used to calculate center-level averages included more than 13,000 youth with complete NJCS baseline data who were predicted to attend the 100 centers in our primary sample.

[4]Of the 100 centers in our sample, 44 centers had between 10 and 100 treatment and control group center designees, 42 centers had between 100 and 200, and 14 centers had more than 200.

After generating estimates of participant characteristics for each center, we linked these characteristics to the Job Corps performance measure data. We then constructed regression-adjusted performance measures using similar procedures to those used to adjust the JTPA performance standards in the 1980s and 1990s (see Dickinson et al., 1988; Social Policy Research Associates, 1999) [**Step 3 of Figure 1**]. Specifically, we estimated the following regression model:

(1) $PM_c = X_c\beta + \varepsilon_c$,

where PM_c is the center's performance measure, X_c is a row vector of center-level baseline participant characteristics (including the intercept), and β is the parameter vector to be estimated. The mean-zero error, ε_c, is the component of the center's performance level that cannot be explained by X_c.

Equation (1) was estimated using ordinary least squares, where each center was weighted equally (although in our sensitivity analysis we also estimated models giving more weight to the larger centers). The estimated residual,

(2) $\hat{\varepsilon}_c = PM_c - X_c\hat{\beta}$

is the regression-adjusted performance measure for center c, where $\hat{\beta}$ is the estimated parameter vector. This residual represents the part of the center's performance level that is not due to the types of students served or local area economic conditions that are measured in X_c. We obtained very similar results using an alternative model that relied on individual-level data where PM_c was regressed on center-level indicators and individual-level X variables, and where the estimated parameters on the center-level indicators were used as the regression-adjusted performance measures (not shown).

We used various specifications for PM_c and X_c in equation (1). For PM_c, we used the center's overall performance rating, as well as components of that rating. In addition to using a two- or three-year average of performance measures, we conducted the analysis separately for each PY

between 1994 and 1996. As discussed, we also estimated separate models using X_c variables from the program intake forms and the more comprehensive NJCS baseline survey. Because the NJCS baseline data provide information on more characteristics than we have Job Corps centers, we relied on stepwise regression procedures (using a 0.20 p-value criterion for variable inclusion) to identify the set of baseline characteristics that have the most explanatory power in the model.

Defining Impact Estimates at the Center Level

We generated center-level impact estimates using the center predictions of intake counselors discussed above. The intent-to-treat (ITT) parameter for a center—the effect of being offered the opportunity to attend a given Job Corps center—was estimated as the difference in weighted mean outcomes between treatment and control group members who were designated for that center [**Step 4 of Figure 1**]. Because the treatment group Job Corps center enrollment rate was 73 percent, we instead relied on an instrumental variables approach to estimate impact estimates for *participants*—that is, the treatment-on-the-treated (TOT) parameter—by dividing the ITT estimates by the difference between the treatment group Job Corps enrollment rate and the control group crossover rate (Angrist, Imbens, & Rubin, 1996; Bloom, 1984; Heckman, Smith, & Taber, 1998). In particular, the TOT impact for a center was estimated as:

$$(3)\quad impact_c = \frac{\bar{y}_{treatment,c} - \bar{y}_{control,c}}{p_{treatment,c} - p_{control,c}},$$

where $impact_c$ is the TOT impact estimate for center c, $\bar{y}_{treatment,c}$ is the mean outcome among treatments predicted to attend center c, $\bar{y}_{control,c}$ is the mean outcome among controls, $p_{treatment,c}$ is the Job Corps participation rate among treatments, and $p_{control,c}$ is the Job Corps participation rate

among controls.[5] The means and participation rates were weighted to account for sampling probabilities and survey nonresponse. We calculated center-level impacts for all seven outcomes in Table 5.

In the final step of our analysis, we calculated the correlation between center-level impact estimates (calculated using Equation 3) and the adjusted and unadjusted performance measures [**Step 5 of Figure 1**].

BACKGROUND DESCRIPTIVE ANALYSES

This section presents results from several descriptive analyses to assess whether regression adjustment has the potential to change the unadjusted performance rankings.

Relationship Between Unadjusted Performance Measures and Participant Characteristics

Understanding the relationship between unadjusted performance measures and center-level baseline participant characteristics is critical for determining (1) the extent to which centers with different ratings served systematically different participants, and hence (2) the scope for regression adjustment to influence performance measures. To look at this relationship, we categorized centers into three terciles (with low, medium, and high three-year average unadjusted overall ratings) and then tabulated, for each group, average selected baseline participant and local area characteristics from the NJCS baseline data and the ARF (Table 6).

There are some statistically significant differences in the baseline characteristics of participants across centers with low, medium, and high overall performance (Table 6). In general, participants in high-performing centers had characteristics that are associated with favorable outcomes, with a

[5] Between random assignment and the 48-month follow-up interview, about 73 percent of the treatment group, and 1 percent of the control group, had enrolled in Job Corps.

handful of exceptions. For example, relative to lower-performing centers, high-performing centers had a larger share of students who were white, with a high school degree, and who were from areas with higher incomes. However, students at high-performing centers were more likely to have used drugs and less likely to be native English-speakers. Furthermore, the differences across groups are relatively small, and there are several student characteristics on which centers in different terciles do not differ, including gender, age, and arrests. The pattern of results is similar using the program intake data and the full set of NJCS baseline survey data items (not shown).

Based on these findings, it is unclear how the relative rankings of different centers will change due to the regression adjustment of performance measures. However, there is some scope for change, given the differences in characteristics across center performance categories.

Correlations Between Different Performance Measures

To understand the relationship between the different performance measure components, we estimated Spearman rank correlations between center rankings that were constructed using different components (Table 7). The center rankings are, in general, positively correlated, and tend to be higher for measures within the same area (program achievement and placement) than across areas. There is, however, considerable heterogeneity in the rank correlations, and the pairwise correlations are not universally large. Because it appears that different performance measures are capturing different dimensions of center performance, our analysis uses the overall performance measure used by Schochet and Burghardt (2008) as well as each component measure.

Distributions of Performance Measures and Estimated Impacts Across Centers

An important issue for assessing the extent to which the regression adjustment of the performance measures can make a difference is to examine variation in both performance measures and estimated impacts across centers. We find that the multiyear average performance measures do

not vary substantially across the 100 study centers (Table 8). Median values range from 1 to 1.26, and the measures range from 0.56 to 2.19. There are, however, some differences across measures. For instance, the placement rate, average wage, full-time placement, and ARPA rating measures vary less than the other measures, and the reading gains, math gains, and GED rate measures vary the most. These results suggest that small changes in the performance measures due to regression adjustment could have a nontrivial effect on the center performance rankings.

The center-level impact estimates vary considerably more than do the performance measures (Table 8). For instance, impacts on 1998 annual earnings range from -$8,566 to $10,908, and impacts on the receipt of a vocational certificate range from -20.3 to 69.7 percentage points. These results could be partly due to relatively small sample sizes in some centers. Thus, in some of our analyses below, we group centers into performance terciles to help reduce measurement error.

COMPARING ADJUSTED AND UNADJUSTED PERFORMANCE MEASURES

This section presents evidence that although the regression adjustment process changes somewhat the center performance measures and rankings, there remains a positive relationship between the unadjusted and adjusted measures.

R^2 Values From the Regression Adjustment Models

As described above, we estimated the relationship between unadjusted performance measures and center-level averages of participant characteristics using equation (1). We tested the sensitivity of our findings to the inclusion of different covariates, including NJCS baseline characteristics, program intake data, local area characteristics, and center characteristics. Table 9 shows R^2 values from a representative set of models that were used to adjust the performance measures. We show

results for five different performance measures (all three-year averages) and various specifications for the covariates included in the models.

The results indicate that the baseline characteristics have significant explanatory power in the models—in other words, that the characteristics of participants, their local areas, and centers are correlated with performance measures—although the R^2 values vary somewhat across performance measures and specifications (Table 9). Depending on the model and performance measure, between 41 percent and 89 percent of the variance is explained by the covariates. These R^2 values are higher than those from the JTPA adjustment models, which were typically less than 0.2 (Barnow & Smith, 2004). This suggests that there is scope for the regression adjustment process to influence center performance rankings.

An unexpected finding is that the R^2 values tend to be slightly larger using the program intake data than using the NJCS data. This could be due to the stepwise regression procedure that was used to select the model covariates in specifications that used the NJCS data; this process was not used in specifications using the program intake data where all variables were included in the models. In fact, more program intake variables were included in the models than NJCS baseline variables. Adjusted R^2 values (that account for the number of model covariates) are larger using the NJCS variables (not shown). It is striking, however, that the two data sources have similar predictive power.

Correspondence Between the Unadjusted and Adjusted Performance Measures

To get a sense for how regression-adjusted performance measures relate to the unadjusted measures, we identified low, medium, and high performers according to each measure and used these groupings to construct three-by-three contingency tables (Table 10 shows an example for the three-year average overall rating). We then examined cell counts in the diagonal and off-diagonal entries.

Using this approach, we find that regression adjustment has some influence on center performance rankings. For instance, using the three-year average overall rating, nearly half of all centers are classified into new terciles after adjustment (Table 10). Using the NJCS adjustment, 52 centers have adjusted and unadjusted three-year average overall ratings that are in the same tercile, and 48 have ratings in different terciles, and similarly for the program intake adjustment. The share of centers with matching terciles is similar for other performance measures as well (not shown).

In addition to comparing terciles of the performance measure distributions, we calculated the correlations between unadjusted and adjusted performance measures (Table 11). These correlations range from 0.39 to 0.82, depending on the measure, year, and adjustment. Surprisingly, correlations are generally higher using the NJCS adjustment than the program intake adjustment. However, as discussed, this difference could reflect the use of the stepwise procedure for the NJCS adjustment, rather than the data source.

Finally, Figure 2 presents scatter plots showing the relationship between unadjusted and adjusted performance measures for the three-year average overall rating. The figure also shows the fitted regression lines through these points. Consistent with the results above, the plots show that the unadjusted and adjusted performance measures are positively correlated. However, there is some scatter around these regression lines, which indicates that performance measure rankings were, to some degree, changed by adjustment.

COMPARING PERFORMANCE MEASURES AND IMPACT ESTIMATES

To examine the extent to which the regression adjustment process improved the performance-impact association, we compared the adjusted and unadjusted performance measures ($\hat{\varepsilon}_c$ and PM_c) to the center-level impact estimates ($impact_c$). We summarize these relationships in Table 12 which presents correlation coefficients between key performance measures and impacts.

We find that the performance-impact correlation coefficients are generally small and statistically insignificant for both the adjusted and unadjusted performance measures. The notable exception is the negative correlation between several performance measures and the 1997 and 1998 earnings impacts. Figure 3 further demonstrates this relationship by showing a plot of the overall performance measure against the 1998 earnings impacts for each center.

Another approach for assessing the relationship between the performance measures and impacts is to separately group centers into terciles based on their performance rankings and impacts and to create a contingency table that crosses these two groupings. Table 13 presents a representative contingency table using the 1998 earnings impacts and the adjusted overall performance rankings. The table clearly indicates that there is little association between the impact tercile and the overall rating tercile for the adjusted performance measures. Statistical chi-square tests confirm the independence of the impact and performance tercile counts. The associations are even weaker using the other performance measures.

Finally, we conducted myriad sensitivity analyses to examine the robustness of our core finding of no relationship between unadjusted and adjusted performance measures and impacts. For instance, we examined the relationship using performance measures for each PY between 1994 and 1996, and conducted analyses using regression adjustment models that included center characteristics. In addition, to help adjust for measurement error in the center-level impact estimates, we conducted analyses where we restricted the sample to centers with 100 or more NJCS follow-up observations and estimated models where centers were weighted proportional to their size rather than equally. In all specifications, we found that the adjusted performance ratings generally remain uncorrelated with center-level impacts.

DISCUSSION

Our finding of zero relationship between performance measures and impacts is consistent with Barnow (2000) and Heckman, Heinrich, and Smith (2002). Furthermore, our results are consistent with previous studies that have found that impact estimates based on nonexperimental methods do not mimic those based on experimental methods (see, for example, Fraker & Maynard, 1987; LaLonde, 1986; Peikes et al., 2008; Smith & Todd, 2005; Glazerman, Levy, & Myers, 2003). Here, regression-adjusted performance measures—which are analogous to nonexperimental impact estimates—are not strongly associated with impact estimates from the experimental NJCS study. While the baseline covariates did explain some of the variance in the performance measures, the lack of association between adjusted performance measures and impacts suggests that there are likely to be important unobserved differences between students attending different centers and that those unobserved factors may be associated with outcomes.

It is possible that measurement error may be influencing our results. First, the impact estimates may be imprecisely estimated due to relatively small sample sizes in some of the 100 centers. However, our analyses that grouped centers into performance terciles helped adjust for this imprecision and corroborated our overall correlational analyses. Another potential source of measurement error stems from the possible mismatch of the NJCS survey data that was used to construct the impact estimates and the Job Corps performance data. As shown in Table 14, correlations between the performance measures and treatment group participant outcomes averaged to the center level are only moderately correlated. This could reflect differences in the samples used to construct the measures, in the time frame over which data are collected, or in how outcomes are defined and reported. If center-level treatment group participant outcomes are used as the

"performance measures," the performance-impact correlations are larger and sometimes significant (not shown).

The weak association between performance measures and impacts could also be related to the fact that performance measures do not vary a great deal across centers. For example, the three-year average overall Job Corps performance measure varies from 0.87 to 1.34, and the interquartile range is only 0.10 points. Though this implies that there may be scope for regression adjustment to reorder centers, it has the drawback that regression adjustment may actually highlight small differences between center performance measures that may be not meaningful.

CONCLUSIONS

USDOL has used performance management systems for its major workforce programs since the 1970s. Yet there is only limited evidence on the extent to which performance measures for these programs—and in particular, adjusted performance measures that "level the playing field"—track estimates of causal program effects in improving participants' outcomes relative to the counterfactual of what they would have been in the absence of the program.

This article has contributed to the literature by examining the extent to which adjusted performance measures yield accurate assessments of program impacts using data from a large-scale, experimental evaluation of Job Corps. Job Corps is a pertinent case study for the analysis because it has used a rigorous and widely-emulated performance standards system since the late 1970s and is a performance-driven program.

At the time of the NJCS, Job Corps minimally adjusted only a small number of its performance measures. Thus, we used extensive NJCS baseline survey, program intake, and local area data to fully adjust the Job Corps performance measures (for different components and different program years) and to examine their correlation with center-level impact estimates for a range of outcomes

(educational services, educational attainment, arrests, and earnings). The estimation of these center-level impacts was complicated by the lack of information on actual center assignments for control group members as well as for treatment group members who did not enroll in the program. To overcome this complication, we used a novel approach, where center-level impacts were estimated using accurate predictions of program intake staff, made at the time of program application, about the centers to which all treatment and control applicants would likely be assigned.

We found that students in high-performing centers were somewhat more likely to have characteristics associated with positive labor market outcomes than students in lower-performing centers, although the differences are relatively small. We found also that the participant, local area, and center characteristics used in the regression adjustment models typically explained about 65 percent of the variance of the performance measures, which led to some differences in the adjusted and unadjusted performance measures (correlations between these two measures were typically about 0.58). However, we found that the regression-adjusted performance measures were no better than the unadjusted performance measures at distinguishing between centers with larger impacts and those with smaller impacts. The performance-impact correlations were generally small and statistically insignificant for both the overall measures of performance as well as each component of center performance. Similar results held whether using program intake data or the more detailed NJCS baseline survey data for regression adjustment.

In sum, while the detailed baseline covariates did explain some of the variance in the performance measures, there appear to remain important unobserved differences between students attending different centers that are correlated with key participant outcomes. Our results suggest that adjusted performance measures might not serve as a substitute for impact estimates.

REFERENCES

Angrist, J., Imbens, G., & Rubin, D. (1996). Identification of causal effects using instrumental variables. Journal of the American Statistical Association, 91, 444-455.

Barnow, B. (2000). Exploring the relationship between performance management and program impact: A case study of the Job Training Partnership Act. Journal of Policy Analysis and Management, 19, 118-141.

Barnow, B. (2006). Performance measures adjustment and incentives: key strategies for providing improved services to harder to serve populations in the age of accountability. Washington DC: Report for the National Collaborative on Workforce and Disability.

Barnow, B. (2009). The role of performance management in workforce investment programs. College Park, MD: Working Paper, University of Maryland, School of Public Policy.

Barnow, B. & Heinrich, C. (2010). One standard fits all? The pros and cons of performance standard adjustments. Public Administration Review, 70, 60-71.

Barnow, B. & Smith, J. (2004). Performance management of U.S. job training programs. In C. O'Leary, R. Straits, and S. Wandner (Eds.), Job training policy in the United States. Kalamazoo, MI: The W.E. Upjohn Institute for Employment Research.

Bartik, T. (1995). Using performance indicators to improve the effectiveness of welfare-to-work programs. Kalamazoo, MI: Upjohn Institute Working Paper No. 95-36.

Blalock, A. & Barnow, B. (2001). Is the obsession with performance management masking the truth about social programs? In D. Forsyhte (Ed.), Managing performance in American government. Rockefeller Institute Press.

Bloom, H. (1984). Accounting for no-shows in experimental evaluation designs. Evaluation Review, 8, 225-246.

Courty, P., Heinrich, C., Marschke, G., & Smith, J. (2011). U.S. employment and training programs and performance standards system design. In J. Heckman, C. Heinrich, P. Courty, G. Marschke, and J.Smith (Eds.), The performance of performance standards. Kalamazoo, MI: The W.E. Upjohn Institute for Employment Research, Kalamazoo.

Dickinson, K., West, R., Kogan, D., Drury, D., Franks, M., Schlichtmann, L., & Vencill, M. (1988) Evaluation of the effects of JTPA performance standards on clients, services and costs. Washington DC: National Commission for Employment Policy, Research Report 88-16.

Fortson, J. & Schochet, P.Z. (2011). Analysis of associations between contemporaneous Job Corps performance measures and impact estimates from the National Job Corps Study. Washington DC: Final Report Submitted to the U.S. Department of Labor.

Fraker, T. & Maynard, R. (1987). The adequacy of comparison group designs for evaluations of employment-related programs. Journal of Human Resources, 22.

Glazerman, S., Levy, D., and Myers, D. (2003). Nonexperimental versus experimental estimates of earnings impacts. The Annals of the American Academy of Political and Social Science, 589, 63-93.

Heckman, J., Heinrich, C. & Smith, J. (2002). The performance of performance standards. Journal of Human Resources, 37, 778-811.

Heckman, J. & Smith, J. (2011). Do the determinants of program participation data provide evidence of cream skimming? In J. Heckman, C. Heinrich, P. Courty, G. Marschke, and J.Smith (Eds.), The performance of performance standards. Kalamazoo, MI: The W.E. Upjohn Institute for Employment Research, Kalamazoo.

Heckman, J., Smith, J., & Taber, C. (1998). Accounting for dropouts in evaluations of social programs. Review of Economics and Statistics, 130, 1-14.

Heinrich, C. (2004). Improving public sector performance management: One step forward, two steps back? Public Finance and Management, 4, 317-351.

Hill, C. (2003). Impacts, outcomes, and management in welfare-to-work programs. Georgetown Public Policy Institute Working Paper.

LaLonde, R. (1986). Evaluating the econometric evaluations of training programs with experimental data, American Economic Review, 76, 604-620.

Orr, L., Bloom, H., Bell, S., Doolittle, F., Lin, W., & Cave, G. (1996). Does training for the disadvantaged work? Evidence from the National JTPA Study. Washington, DC: Urban Institute Press.

Peikes, D., Moreno, L. & Orzol, S.M. (2008). Propensity score matching: A note of caution for evaluators of social programs, The American Statistician, 62, 222-231.

Schochet, P.Z. (2001). National Job Corps Study: Methodological appendixes on the impact analysis. Princeton, NJ: Mathematica Policy Research, Inc.

Schochet, P.Z. & Burghardt, J. (2008). Do Job Corps performance measures track program impacts? Journal of Policy Analysis and Management, 27, 556-576.

Schochet, P.Z., Burghardt, J., & McConnell, S. (2008). Does Job Corps work? Impact findings from the National Job Corps Study. American Economic Review, 68, 1864-1886.

Siedlecki, J. & King, C. (2005). Approaches to adjusting workforce development performance measures. Occasional Policy Brief, 1, Lyndon B. Johnson School of Public Affairs, University of Texas, Austin.

Smith, J. & Todd, P. (2005). Does matching overcome LaLonde's critique of nonexperimental estimators? Journal of Econometrics, 125, 305-53.

Social Policy Research Associates (1999). Guide to JTPA performance standards for program years 1998 and 1999. Oakland, CA: Report submitted to the U.S. Department of Labor, Employment and Training Administration.

U.S. Department of Health and Human Services, Health Resources and Services Administration, Area Resource File (ARF). (2008). Rockville, MD.

Table 1. Job Corps Center Performance Measures, PYs 1994-1996

Measure	Years Available		
	PY 1994	PY 1995	PY 1996
Reading Gains. Percentage of students who gain two grades or score above a threshold on follow-up TABE reading test (among those who did not take or scored less than the threshold on TABE 5/6 total reading test at program entry).	X	X	
Math Gains. Percentage of students who gain two grades or score above a threshold on follow-up TABE math test (among those who did not take or scored less than the threshold on TABE 5/6 total math test at program entry).	X	X	
GED Rate. Percentage of students who obtain GED/high school degree, including bonus for students who initially score low on test (among those without high school diploma and who either did not take or scored more than a threshold on TABE 5/6 or TABE 7/8, depending on the year, total reading test at program entry).	X	X	X
Vocational Completion Rate. Percentage of students who complete vocational training at completer or advanced-completer level (depending on the year, either among all terminees or among those who stayed at least 60 days and participated in a vocational program with an approved training achievement record).	X	X	X
Placement Rate. Percentage of students placed in job/military or school, with bonus for advanced training (AT) or advanced career training (ACT) transfers (among terminees and Job Corps AT or ACT transfers).	X	X	X
Average Wage. Average wage of students placed in a job/military.	X	X	X
Quality Placement. Percentage placed in a job training match, with or without a bonus for students placed in college or AT/ACT transfers (depending on the year). Measured among either all job/military completers or vocational completers with a placement record and those with a record that was due but not received (depending on the year).	X	X	X
Full Time. Percentage of students placed who are full-time among students placed in a job/military.		X	X
ARPA Rating. Regional office rating of center quality/compliance.	X	X	
Overall Rating. Weighted average of individual ratings (measure/standard).	X	X	X

Notes: "PYs 1994-1996" refers to PY 1994 (July 1, 1994 to June 30, 1995), PY 1995 (July 1, 1995 to June 30, 1996), and PY 1996 (July 1, 1996 to June 30, 1997). Schochet and Burghardt (2008) provide additional detail on performance measures and the weights used in constructing the overall rating.

Table 2. Measures of Baseline Characteristics from the National Job Corps Study

Measure
Demographic Characteristics. Race, gender, age, native language*, geographic region, local area population density*.
Education and Skills. High school degree*, GED*, vocational degree*, highest grade completed, months in school in past year*.
Employment History. Ever worked*, job in past year*, currently working, months worked in past year*, occupational category of most recent job*, earnings in past year*, physical or emotional problem that limited work*.
Family Status. Marital status*, has child, pregnant*.
Socioeconomic Status. Receipt of welfare in childhood*, receipt of AFDC in past year, receipt of food stamps in past year, currently in public housing*.
Criminal History. Ever arrested*.
Drug Use. Frequency of marijuana use in past year*, frequency of use of hard drugs in past year*, ever in drug treatment*.

* Indicates measures that are not included in the program intake form.

Table 3. Measures of Baseline Characteristics from the Program Intake Form

Measure
Demographic Characteristics. Race, gender, age, city size, prior military service, legal US resident.
Education and Skills. Months out of school, highest grade completed.
Employment History. Weeks since employed full-time, earnings per hour, annual income.
Family Status. Family status (head, related, etc.), number of dependents, needs child care.
Socioeconomic Status. Family in receipt of public assistance.
Criminal History. Convicted or adjudged delinquent.
Health and Health Care. Serious illness, under doctor's care, being treated, health insurance coverage.

Table 4. Measures of Local Area Characteristics from the Area Resource File

Measure of Population	Year	Source
Demographic Characteristics		
Percentage white	1990	1990 Census STF1A
Percentage black	1990	1990 Census STF1A
Average household size	1990	1990 Census STF1A
Percentage urban	1990	1990 Census STF3A
Percentage of families with a female head	1990	1990 Census STF1A
Percentage foreign-born	1990	County and City Data Book, 1994
Total births	1995	Census Estimates of Population
Percentage of births to teens <18 years	1998-2000 1995	1998-2000 NCHS Natality Tape, 1995 Census Estimates of Population
Crime		
Deaths by homicide and legal intervention (rate)	1998-2000 1995	1998-2000 NCHS Mortality Tape, 1995 Census State and County Population Estimates Components of Change
Percentage in juvenile institutions	1990	1990 Census STF1A
Economic Characteristics		
Percentage of families in poverty	1989	1990 Census STF3A
Median household income	1995	Census Small Area Income and Poverty Estimates
Percentage of households in different income categories (<$5,000, $5,000-$9,999, $10,000-$14,999, $15,000-$24,999, $25,000-$49,999, $50,000-$99,999, >$100,000)	1989	1990 Census STF3A
Unemployment rate, 16+	1995	Bureau of Labor Statistics

Notes: STF1A = Census of Population and Household Summary Tape File 1A; STF3A = Census of Population and Housing Summary Tape File 3A; Census Estimates of Population = Population of Counties and Demographic Components of Population Change Time Series, U.S. Bureau of the Census; NCHS = National Center for Health Statistics (Centers for Disease Control and Prevention). Percentage of births to teens < 18 years, deaths by homicide and legal intervention (rate), percentage of population in juvenile institutions, and average household size are calculated using multiple statistics.

Table 5. Outcome Measures from the National Job Corps Study

Measures	Time Frame
Educational Services. Percentage of youth who participated in education and training. Total hours of participation in education and training.	During 48 months after random assignment
Educational Attainment. Percentage of youth who received a GED among those without a high school credential at baseline. Percentage of youth who received a vocational certificate.	During 48 months after random assignment
Arrests. Percentage of youth who were ever arrested.	During 48 months after random assignment
Earnings. Annual earnings based on the survey data.	1997, 1998

Figure 1. Design of Analysis of Job Corps Performance Measures and Impact Estimates

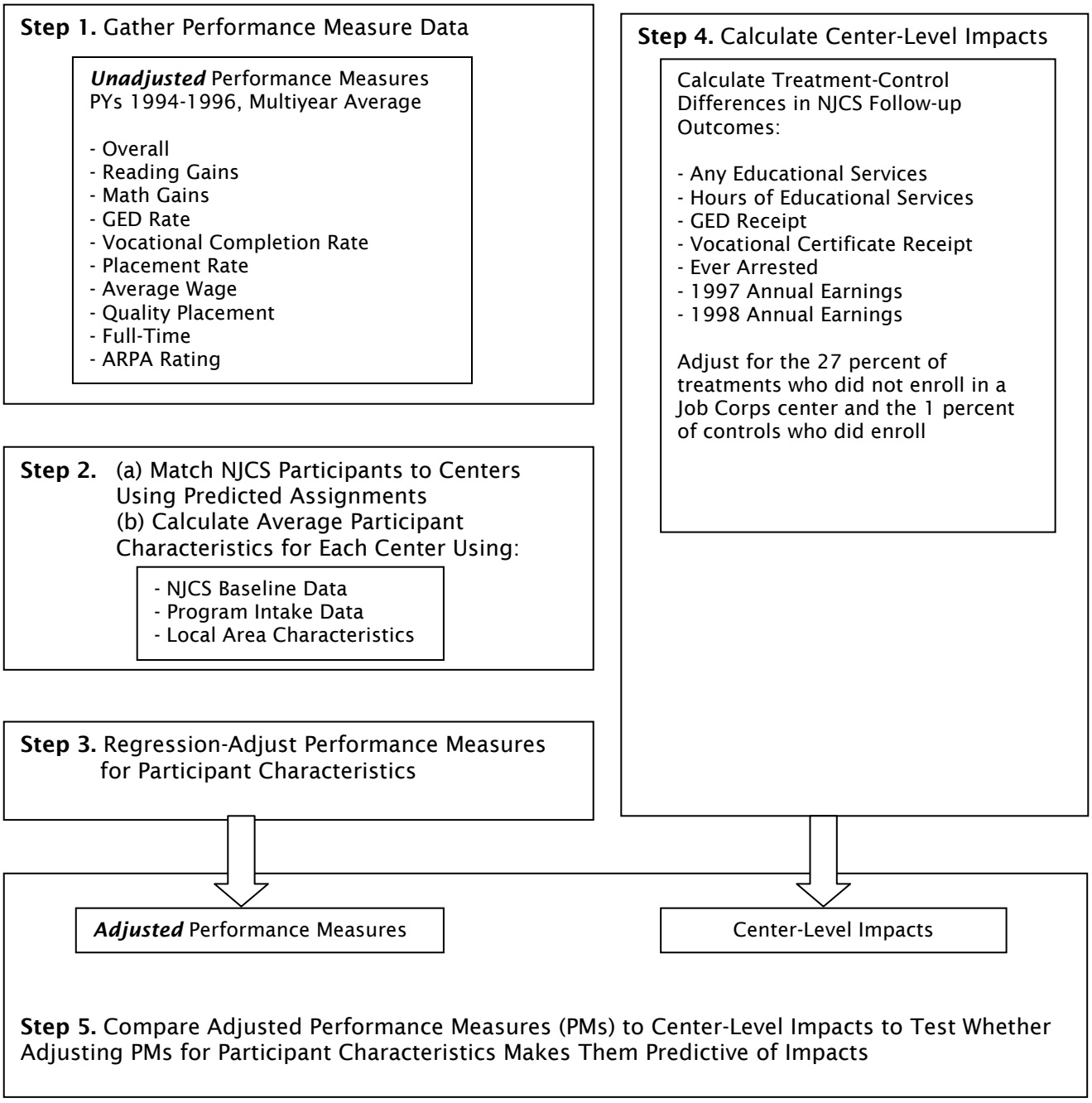


Table 6. Selected Average Baseline Characteristics, by Overall Center Performance Tercile

	Unadjusted Overall Center Performance Tercile			
	Low	Medium	High	*p-value*
Selected NJCS Baseline Survey Characteristics				
Demographic Characteristics				
Non-Hispanic White	0.245	0.344	0.391	0.023*
Non-Hispanic Black	0.558	0.479	0.249	0.000*
Hispanic	0.139	0.106	0.227	0.017*
Other race	0.058	0.071	0.133	0.040*
Female	0.388	0.361	0.399	0.663
Native language English	0.897	0.916	0.799	0.003*
Age 15-17	0.443	0.447	0.427	0.708
Age 18-20	0.395	0.401	0.415	0.498
Age >20	0.162	0.153	0.158	0.817
Education and Skills				
High school degree	0.151	0.152	0.191	0.034*
GED	0.042	0.047	0.057	0.147
Vocational degree	0.025	0.024	0.029	0.465
Employment History				
Currently working	0.202	0.205	0.228	0.203
Earnings in past year <$1,000	0.507	0.483	0.497	0.590
Physical or emotional problem that limited work	0.054	0.050	0.058	0.517
Family Status				
Has child	0.193	0.168	0.148	0.142
Socioeconomic Status				
Did not receive food stamps over past year	0.574	0.602	0.630	0.088
Criminal History and Drug Use				
Ever arrested	0.267	0.298	0.301	0.231
Used no drugs over past year	0.711	0.663	0.622	0.001*
Local Area Characteristics				
Demographic Characteristics				
Percentage white	0.708	0.740	0.790	0.014*
Percentage black	0.221	0.195	0.093	0.000*
Average household size	2.723	2.683	2.789	0.037*
Percentage urban	0.750	0.699	0.772	0.127
Percentage of families with a female head	0.199	0.192	0.171	0.020*
Percentage foreign-born	0.621	0.563	0.835	0.226
Percentage of births to teens <18 years	0.057	0.051	0.049	0.032*
Crime				
Deaths by homicide and legal intervention (rate)	0.000	0.000	0.000	0.012*
Percentage of population in juvenile institutions	0.000	0.000	0.001	0.023*
Economic Characteristics				
Percentage of families in poverty	0.140	0.121	0.121	0.129
Median household income	31726	33230	34064	0.049*
Unemployment rate, 16+	0.061	0.060	0.067	0.117
Number of Centers	**33**	**33**	**34**	

Sources: Performance measure data, NJCS baseline survey and 2008 ARF.

Notes: All centers are weighted equally; when constructing center-level averages, baseline characteristics are weighted using the baseline weight. Terciles are based on the three-year average overall rating. The reported p-value refers to an F-test which tests whether the three groups are jointly significant.

* Statistically significant at the 5 percent level.

Table 7. Rank Correlations Between Unadjusted Center Performance Components, Multiyear Averages

	Center Ranking									
Center Ranking	Overall	Reading Gains	Math Gains	GED	Vocational Completion	Placement	Average Wage	Quality Placement	Full-Time	ARPA Rating
Overall	1.00									
Reading Gains	0.69	1.00								
Math Gains	0.76	0.84	1.00							
GED Rate	0.66	0.48	0.53	1.00						
Vocational Completion Rate	0.87	0.53	0.62	0.52	1.00					
Placement Rate	0.61	0.31	0.40	0.29	0.36	1.00				
Average Wage	0.44	0.08	0.21	0.19	0.36	0.36	1.00			
Quality Placement	0.57	0.23	0.31	0.17	0.39	0.46	0.56	1.00		
Full-Time	-0.07	-0.09	-0.09	-0.35	-0.01	-0.08	0.20	0.19	1.00	
ARPA Rating	0.81	0.60	0.63	0.54	0.60	0.53	0.19	0.44	-0.13	1.00

Sample Size = 100 centers

Source: Performance measure data.

Notes: All centers are weighted equally. Table shows the Spearman rank correlation based on a multiyear average of the center's performance ratio. For reading gains, math gains, full-time, and ARPA rating, this is a two-year average (see Table 1 for available years). For other performance measures, this is a three-year average (PYs 1994-1996).

Table 8. Summary Statistics for Unadjusted Performance Measures and Center-Level Impact Estimates

	Min	1st Quartile	Mean	Median	3rd Quartile	Max	Standard Deviation
Performance Measures (Multiyear Averages)							
Overall	0.87	1.04	1.09	1.10	1.14	1.34	0.09
Reading Gains	0.56	0.96	1.15	1.13	1.33	1.85	0.26
Math Gains	0.59	1.06	1.20	1.20	1.35	1.87	0.24
GED Rate	0.62	0.92	1.06	1.04	1.14	2.19	0.25
Vocational Completion Rate	0.67	0.98	1.09	1.10	1.20	1.48	0.16
Placement Rate	0.90	1.04	1.10	1.11	1.16	1.23	0.08
Average Wage	0.89	0.99	1.03	1.02	1.06	1.16	0.06
Quality Placement	0.95	1.18	1.26	1.26	1.36	1.59	0.13
Full-Time	0.93	1.07	1.10	1.09	1.13	1.21	0.05
ARPA Rating	0.80	0.95	0.99	1.00	1.03	1.10	0.07
Center-Level Impact Estimates							
Any Educational Services[a]	-4.6	20.3	30.0	29.0	39.2	73.1	15.7
Hours of Educational Services	-329	617	945	965	1,232	1,849	463
GED Receipt[a]	-23.2	9.1	21.3	20.2	32.5	115.8	19.5
Vocational Certificate Receipt[a]	-20.3	23.0	31.6	32.1	41.8	69.7	15.4
Arrested[a]	-60.2	-17.2	-6.5	-5.2	3.8	38.9	17.5
1997 Annual Earnings[b]	-8,274	-1,130	494	460	2,607	7,205	3128
1998 Annual Earnings[b]	-8,566	-601	1,415	1,307	3,774	10,908	3448

Sample Size = 100 centers.

Sources: Performance measure data, NJCS follow-up surveys.

[a] Impacts are measured in percentage points.

[b] Impacts are measured in 1995 dollars.

Table 9. Regression R^2 Values from Regressions of Three-Year Average Unadjusted Center Performance Ratings on Center-Level Baseline Characteristics

	Regression R^2				
	Dependent Variable: Unadjusted Performance Rating				
Independent Variables	Overall	GED	Vocational Completion	Average Wage	Placement Rate
NJCS Baseline Characteristics and Local Area Characteristics	0.66	0.76	0.41	0.79	0.70
NJCS Baseline Characteristics, Local Area Characteristics, and Center Characteristics	0.58	0.65	0.59	0.79	0.76
Program Intake Baseline Characteristics and Local Area Characteristics	0.72	0.80	0.58	0.84	0.76
Program Intake Baseline Characteristics, Local Area Characteristics, and Center Characteristics	0.81	0.86	0.71	0.89	0.81

Sample size = 100 centers.

Sources: Performance measure data, NJCS baseline survey, program intake form, center characteristics, 2008 ARF.

Notes: All centers are weighted equally; when constructing center-level averages, baseline characteristics are weighted using the baseline weight. Table reports R^2 values from regressions of three-year average unadjusted performance measures on center-level average baseline characteristics. The regressions that control for NJCS baseline characteristics are forward-selection stepwise regressions with inclusion and exclusion p-value thresholds of 0.20.

Table 10. Unadjusted Center Performance Terciles and Adjusted Center Performance Terciles, Overall Three-Year Average Rating

	NJCS-Adjusted Performance Tercile			Intake-Adjusted Performance Tercile		
Unadjusted Performance Tercile	Low	Medium	High	Low	Medium	High
Low	19	10	4	22	8	3
Medium	10	13	10	6	12	15
High	4	10	20	5	13	16
Number of Centers	33	33	34	33	33	34

Sample size = 100 centers.

Sources: Performance measure data, NJCS baseline survey, program intake form, 2008 ARF.

Notes: Table shows terciles of the three-year average overall performance rating. NJCS-adjusted and intake-adjusted performance terciles are terciles based on adjustments that also include local area characteristics (from the 2008 ARF) but not center characteristics.

Table 11. Correlations Between Unadjusted and Adjusted Center Performance Measures, All Components in All Years

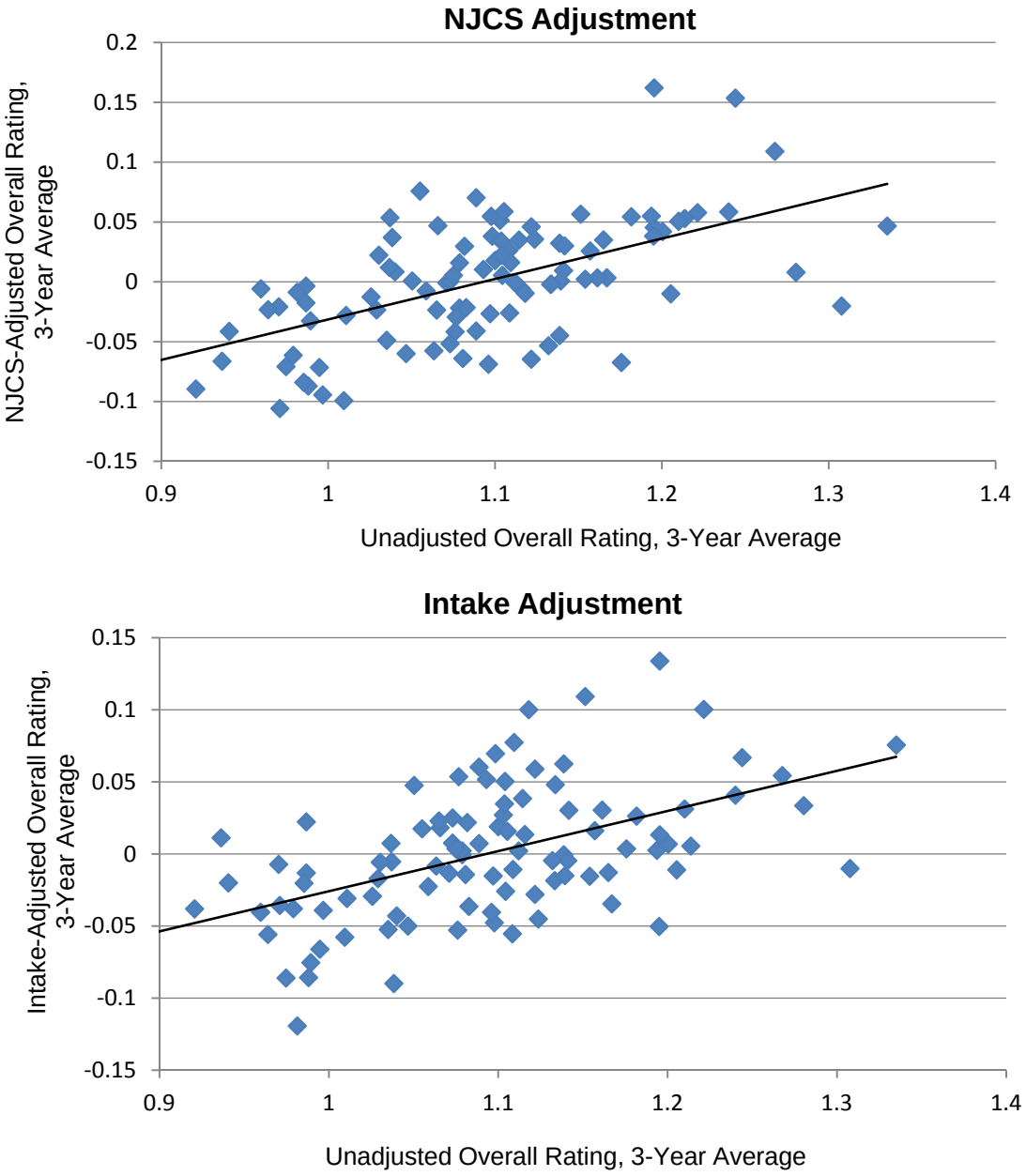
	Correlation Between Unadjusted and Adjusted Performance Measures							
	NJCS-Adjusted				Intake-Adjusted			
Performance Measure	PY94	PY95	PY96	Multiyear Average	PY94	PY95	PY96	Multiyear Average
Overall	0.54	0.69	0.65	0.58	0.48	0.58	0.59	0.53
Reading Gains	0.54	0.74	--	0.74	0.61	0.67	--	0.64
Math Gains	0.66	0.82	--	0.75	0.55	0.64	--	0.56
GED Rate	0.53	0.64	0.64	0.49	0.39	0.54	0.57	0.45
Vocational Completion Rate	0.63	0.71	0.71	0.77	0.63	0.65	0.67	0.65
Placement Rate	0.55	0.56	0.51	0.55	0.48	0.51	0.52	0.49
Average Wage	0.59	0.40	0.48	0.46	0.45	0.41	0.48	0.40
Quality Placement	0.74	0.66	0.65	0.63	0.55	0.53	0.54	0.49
Full-Time	--	0.55	0.65	0.56	--	0.50	0.54	0.52
ARPA Rating	0.72	0.62	--	0.66	0.60	0.62	--	0.58

Sample size = 100 centers.

Sources: Performance measure data, NJCS baseline survey, program intake form, 2008 ARF.

Notes: All correlations are statistically significant at the 1 percent level. NJCS-adjusted and intake-adjusted ratings are based on adjustments that also include local area characteristics (from the 2008 ARF) but not center characteristics. All centers are weighted equally; when constructing center-level averages, baseline characteristics are weighted using the baseline weight.

Figure 2. Unadjusted and Adjusted Center Performance Measures, Three-Year Average Overall Rating



Sample size = 100 centers.

Sources: Performance measure data, NJCS baseline survey, program intake form, 2008 ARF.

Notes: NJCS-adjusted and intake-adjusted ratings are based on adjustments that also include local area characteristics (from the 2008 ARF) but not center characteristics. All centers are weighted equally; when constructing center-level averages, baseline characteristics are weighted using the baseline weight. In both graphs, the slopes are statistically significant at the 1 percent level.

Table 12. Correlations Between Center-Level Impacts and Multiyear Average Performance Ratings (Unadjusted, NJCS-Adjusted, and Intake-Adjusted)

	Overall Rating			GED Rating			Vocational Completion Rating			Average Wage Rating			Placement Rating		
Outcome for Impact Estimate	Unadj	NJCS-Adj	Intake-Adj	Unadj	NJCS-Adj	Intake-Adj	Unadj	NJCS-Adj	Intake-Adj	Unadj	NJCS-Adj	Intake-Adj	Unadj	NJCS-Adj	Intake-Adj
Any Educational Services	-0.02	-0.06	-0.01	-0.08	0.08	-0.02	0.08	0.06	0.07	0.09	0.13	0.15	0.05	-0.75	0.04
Hours of Educational Services	0.17	-0.03	0.08	0.02	-0.03	-0.03	0.19	0.13	0.06	-0.01	-0.04	-0.04	0.19	0.16	0.12
GED Receipt	0.15	-0.08	-0.10	0.12	-0.10	-0.11	0.13	0.02	-0.12	0.05	0.10	-0.08	0.13	0.06	-0.10
Vocational Certificate Receipt	0.13	-0.04	-0.01	0.00	-0.06	-0.14	0.14	0.04	0.03	-0.11	0.05	-0.07	0.23*	0.12	0.08
Ever Arrested	-0.02	-0.06	-0.09	0.02	0.07	-0.03	-0.06	-0.10	-0.12	-0.04	-0.08	-0.01	-0.04	-0.04	0.03
1997 Annual Earnings	-0.14	-0.19	-0.22*	-0.22*	-0.32*	-0.25*	-0.08	-0.05	-0.13	0.07	0.03	-0.18	0.03	0.04	-0.01
1998 Annual Earnings	-0.09	-0.11	-0.11	-0.28*	-0.32*	-0.23*	-0.02	-0.01	-0.01	0.05	0.02	-0.18	0.08	0.11	0.03

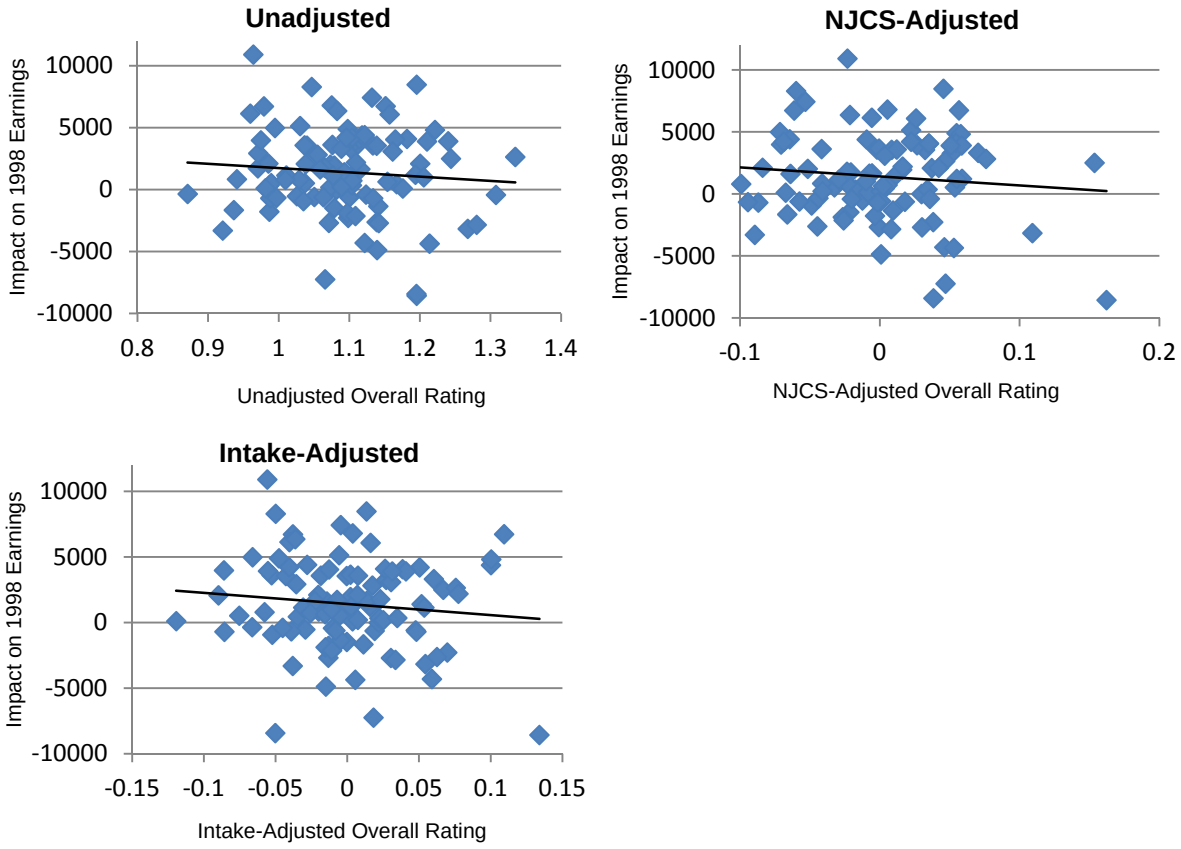
Sample size = 100 centers.

Sources: Performance measure data, NJCS baseline survey, program intake form, 2008 ARF, NJCS follow-up surveys.

Notes: Table shows the correlation based on a multiyear average of the center's performance rating and the center-level impact estimate. NJCS-adjusted and intake-adjusted ratings are based on adjustments that also include local area characteristics (from the 2008 ARF) but not center characteristics. All centers are weighted equally; when constructing center-level averages, baseline characteristics are weighted using the baseline weight. Impacts are calculated using the follow-up weight and are adjusted for differences in participation across research groups.

* Statistically significant at the 5 percent level.

Figure 3. Three-Year Average Overall Center Performance Rating and 1998 Annual Earnings Center-Level Impacts (Unadjusted, NJCS-Adjusted, and Intake-Adjusted Performance)



Sample size = 100 centers.

Sources: Performance measure data, NJCS baseline survey, program intake form, 2008 ARF, NJCS follow-up surveys.

Notes: NJCS-adjusted and intake-adjusted ratings are based on adjustments that also include local area characteristics (from the 2008 ARF) but not center characteristics. All centers are weighted equally; when constructing center-level averages, baseline characteristics are weighted using the baseline weight. Impacts are calculated using the follow-up weight and are adjusted for differences in participation across research groups. In all three graphs, the slopes are not statistically significant.

Table 13. 1998 Annual Earnings Center-Level Impacts and Three-Year Average Overall Adjusted Performance Terciles

	NJCS-Adjusted Three-Year Average Overall Performance Tercile			Intake-Adjusted Three-Year Average Overall Performance Tercile		
1998 Annual Earnings Impacts Tercile	Low	Medium	High	Low	Medium	High
Low	12	11	10	9	12	12
Medium	12	12	9	10	13	10
High	9	10	15	14	8	12
Number of Centers	33	33	34	33	33	34

Sources: Performance measure data, NJCS baseline survey, program intake form, 2008 ARF, NJCS follow-up surveys.

Notes: NJCS-adjusted and intake-adjusted ratings are based on adjustments that also include local area characteristics (from the 2008 ARF) but not center characteristics. All centers are weighted equally; when constructing center-level averages, baseline characteristics are weighted using the baseline weight. Impacts are calculated using the follow-up weight and are adjusted for differences in participation across research groups.

Table 14. Correlations Between Center-Level Treatment Group Outcomes and Unadjusted Performance Ratings

	Unadjusted Performance Rating									
	Overall	Reading Gains	Math Gains	GED	Vocational Completion	Placement	Average Wage	Quality Placement	Full-Time	ARPA Rating
Any Educational Services	0.16	-0.05	-0.02	0.11	0.19	0.07	0.22*	0.10	-0.04	0.12
Hours of Educational Services	0.39*	0.21*	0.20	0.39*	0.29*	0.22*	0.04	0.29*	-0.21*	0.39*
GED Receipt	0.28*	0.08	0.22*	0.21*	0.20*	0.37*	0.26*	0.21*	-0.27*	0.26*
Vocational Certificate Receipt	0.34*	0.17	0.25*	0.13	0.27*	0.29*	0.10	0.32*	-0.13	0.45*
Ever Arrested	-0.13	-0.09	-0.01	-0.29*	-0.04	-0.13	0.22*	0.00	0.20*	-0.19
1997 Annual Earnings	0.02	-0.09	0.03	-0.24*	-0.01	0.31*	0.41*	0.20	0.15	0.02
1998 Annual Earnings	0.11	-0.02	0.05	-0.20*	0.10	0.32*	0.44*	0.25*	0.07	0.08

Sample size = 100 centers.

Sources: NJCS baseline survey, NJCS follow-up surveys.

Notes: Average center-level outcomes for each group (treatment and control) are weighted using the follow-up weight.

* Statistically significant at the 5 percent level.