The Effects of Welfare-to-Work Program Activities on Labor Market Outcomes

Andrew Dyke, *ECONorthwest*

Carolyn J. Heinrich, *University of Wisconsin–Madison*

Peter R. Mueser, *University of Missouri–Columbia and IZA*

Kenneth R. Troske, *University of Kentucky and IZA*

Kyung-Seong Jeon, *University of Missouri–Columbia*

Studies examining welfare-to-work program effectiveness present mixed and sometimes discrepant findings, partly due to research design, data, and methodological limitations. Using administrative data on Missouri and North Carolina welfare recipients, we substantially improve on past estimation approaches to identify the distinct effects of each state's welfare-to-work subprograms—assessment, job search assistance and job readiness training, and more intensive programs designed to augment human capital. More intensive training is associated with greater initial earnings losses but also greater long-run earnings gains. The negative program impacts we observe in quarters

We would like to thank Randy Eberts, Robert LaLonde, Seth Sanders, Jeff Smith, Ed Vytlacil, Jennifer Ward-Batts, and participants at the 2004 Summer Research Workshop at the Institute for Research on Poverty and the 2005 Mark Berger Memorial Conference. This work was supported by a grant from the Rockefeller/Russell Sage Future of Work program. Any errors are our own. Contact the corresponding author, Kenneth R. Troske, at ktroske@uky.edu.

[*Journal of Labor Economics,* 2006, vol. 24, no. 3]

0734-306X/2006/2403-0006$10.00

immediately following participation turn positive by the second year after participation.

I. Introduction

Public expenditures in programs designed to move welfare recipients into the labor market—commonly known as welfare-to-work programs—are made with the expectation that these services will increase participants' employment success both by aiding their search activities and by improving their general skill levels. Yet a growing number of studies examining the effectiveness of state programs to help welfare recipients become employed and exit welfare present mixed and sometimes discrepant findings. Reasons for these mixed findings include inadequacies in research design, data, and methodologies for assessing program effects. Many studies treat welfare-to-work programs as a single entity, when, in fact, these programs typically consist of various subprograms, each providing distinct types of training such as basic education, job preparation and search assistance, and/or vocational and on-the-job training. If outcomes differ by subprogram, then conflicting findings could reflect states' emphases on different subprograms within their overall welfare-to-work programs.

Conflicting findings have, in turn, engendered an active debate over whether strategies designed to provide more intensive training are more effective in assuring welfare recipients' labor market success than those intended to help welfare recipients get into jobs quickly—known as "work-first" approaches. Barnow and Gubits (2002) review a large number of studies of welfare-to-work program outcomes and report that longer-term, intensive training strategies appear to be considerably more effective than short-term work-first strategies. Alternatively, in a meta-analysis of 27 experimental evaluations of 116 welfare-to-work interventions, Greenberg, Cebulla, and Bouchet (2005) determine that job search has persistent positive impacts on labor market outcomes, while basic education, vocational training, and work experience have marginal or negative impacts. Finally, Bloom and Michalopoulos (2001) review experimental evaluations of 20 programs and conclude that the most effective programs combined employment-focused and educational/training strategies with the flexibility to determine the appropriate mix of approaches for a given individual. These conclusions are at least partly at odds with the mid-1990s welfare reforms, which were predicated on the belief that welfare recipients needed jobs, not education and training, to advance in the labor market (Haskins and Blank 2001).

In this article we use administrative data on welfare recipients in the states of Missouri and North Carolina to obtain separate estimates of the effects of participating in subprograms of each state's welfare-to-work program. Our data consist of all women who entered welfare between 1997 Q2 and 1999 Q4. We follow recipients for up to 16 quarters after they enter welfare and model their quarterly earnings as a function of demographic characteristics, prior welfare and work experience, and the specific types of welfare-to-work programs in which they participate.

We divide training into three categories on the basis of the intensity of the activity: participants who went through an assessment but received no other training; participants in job readiness or job search activities; and participants who received more intensive training, including basic education, vocational skills training, or other longer-term programs. We estimate the effect of each training program using propensity score matching techniques that match individuals receiving treatment with similar individuals in a comparison sample. Our results, particularly for Missouri, indicate a clear pattern: over the long run, more intensive training programs produce larger and more persistent returns than short-run work-first strategies. Combined with similar findings in Hotz, Imbens, and Klerman (2006, in this issue), these results suggest that the current emphasis on work-first activities is misplaced and argue for a greater emphasis on training activities designed to enhance participants' human capital.

The remainder of the article is as follows. In the next section we provide a brief review of relevant literature. In Section III, we describe our data and present an overview of our empirical approach. In Section IV, we introduce the various subprograms that are part of each state's welfare-to-work program and describe observed patterns of earnings for participants. Section V spells out the models underlying our estimation strategy, and Section VI gives details of the estimation methods and reports the results of specification tests. Program effect estimates are reported in Section VII, and Section VIII concludes.

II. Evaluation of Welfare-to-Work Programs

Literature reviews by Leahey (2001) and Barnow and Gubits (2002) highlight important areas of disagreement among researchers regarding the effects of programs designed to increase the employment, earnings, and probability of exit among women receiving welfare. Barnow and Gubits (2002) note that one reason for these inconsistent findings is that many studies group low-cost job-oriented activities together with the traditionally more intensive and expensive on-the-job training programs, essentially muddling the distinction between a work-first strategy and one intended to increase general human capital. Some studies are plagued by unreliable records of individuals' participation in specific program activities. Furthermore, as Grogger, Karoly, and Klerman (2002) observe, measures of program characteristics used in many studies, including experimental evaluations, typically only capture variation in official policies, not the actual variation that emerges in the implementation of welfare-to-work programs.

Approaches to aiding welfare recipients that focus on building human capital are rarely fully developed and implemented in welfare-to-work programs (Gueron and Pauly 1991; Jencks and Edin 1992; Friedlander and Burtless 1995). Even among programs that implement such strategies, few have collected sufficient longitudinal data to fully assess their effects. Studies that assess effects beyond 3 years tend to reach different conclusions than short-term assessments; see, for example, both the shorter-term and longer-term analyses of the effects of California's Greater Avenues to Independence (GAIN) program (Riccio, Friedlander, and Freedman 1994; Hotz et al. 2006, in this issue) and Jacobson, LaLonde, and Sullivan's research on displaced workers (1994, 2004).[1]

Definitions of what constitutes longer- versus shorter-term programs also differ widely, complicating comparisons across studies. Some of the "longer-term," intensive education and training strategies in current welfare-to-work programs limit participation to 12 weeks. This contrasts with the earlier Job Opportunities and Basic Skills (JOBS) and Job Training Partnership Act (JTPA) programs, which generally defined long-term training as lasting from 6 months to 2 years. In addition, few studies assess the cumulative effects of multiple short-term episodes of participation in these types of program activities.

More generally, the existing literature makes clear that researchers need better measures of welfare-to-work and training program activities and more precise definitions of what is being measured. Our study aims to improve on the current literature in several ways. We examine participation in specific welfare-to-work program activities over a period when the emphasis on and use of alternative service strategies was changing. With comprehensive information on the types of services provided and timing of participation, we assess the average and cumulative effects of different types of program activities on welfare recipients' outcomes. We use complete data on the populations of welfare recipients in two states, facilitating a comparison of program effects across sites using the same approach and methods of analysis. Finally, we follow participants for up

[1] In the GAIN program, Riverside's emphasis on job search activities contrasted with policies in other counties, particularly Alameda and Los Angeles, which emphasized human capital development and had more registrants in basic skills activities. Adjusted difference-in-difference comparisons of Riverside with Alameda and Riverside with Los Angeles showed that initial differences in program effects (on employment and earnings) were large and in favor of Riverside; however, in later years (4–6), Los Angeles and Alameda had better program outcomes than in Riverside, although the final differences were not statistically significant (Hotz et al. 2006, in this issue).

to 16 quarters after they first enter the program, allowing us to examine the long-run effects of training.

III. Data and Method of Analysis

We examine employment outcomes for welfare recipients in the Temporary Assistance for Needy Families (TANF) programs in Missouri and North Carolina. Our analysis relies on earnings data collected by the states in support of their unemployment insurance programs. Employers report total earnings for each individual in covered employment during each quarter. We merge this information with records used in administering the states' welfare programs, including demographic and household information. While the earnings data omit self-employment, illegal or informal employment, and a small number of jobs not covered by unemployment insurance, the overwhelming majority of employment within each state is included.

For welfare recipients in Missouri, we use employment data collected by the states of Missouri and Kansas, ensuring employment coverage for welfare recipients in Kansas City, Missouri, who often work in Kansas.[2] For welfare recipients in North Carolina, we use that state's employment data. Of course, employment will be understated for individuals who move out of state after leaving welfare.[3] We correct all earnings measures for inflation relative to 1997 Q2. We restrict our sample to female payees, aged at least 18 but less than 65, in single-parent households, excluding "child-only" cases.[4] We use quarters as our time unit, so that an individual who receives TANF cash payments at any point during a given quarter is considered a welfare recipient during that quarter. This approach tends to smooth welfare receipt, eliminating apparent movements off of welfare that are due to administrative errors that cause a case to be omitted from the files for a month or two.[5] We focus on individuals who are new entrants into the TANF cash program during the quarters 1997 Q2–1999 Q4, defining a "new entrant" as one who receives payments during at

[2] About one in seven jobs held by welfare recipients in Jackson County (the central county in the Kansas City metropolitan area) is in Kansas. In St. Louis, the proportion holding jobs in Illinois is much lower.

[3] Kornfeld and Bloom (1999) compare experimental (job-training program) earnings impact estimates calculated using unemployment insurance (UI) earnings data with those based on other more costly earnings data sources and conclude that UI data provide valid estimates for all low-income persons except a small subgroup of male youths with past arrests.

[4] The payee in a child-only case is not a parent and receives payment on behalf of the children. Such payees normally do not face work or training requirements, and their income does not count in the calculation of benefits.

[5] Luks and Brady (2003) studied the definition of welfare spells and concluded that because of "administrative churning," a break of up to 3 months is necessary in most cases to say with confidence that a recipient has gone off of welfare.

least one of these quarters but not the prior quarter. We then follow these individuals for a total of up to 16 quarters, identifying their participation in work component activities and their earnings during each quarter.[6]

Our dependent variable is earnings obtained in a specific outcome quarter. Determinants include individual characteristics, labor market experience and welfare receipt prior to entering welfare, local unemployment rates during the outcome quarter, and work component participation after entering welfare. Since an individual who enters welfare and then obtains adequate employment will subsequently be required to move off of welfare, taking account of welfare exits would be tantamount to controlling for labor market success. We therefore structure our analysis to predict earnings in the 16 quarters beginning with welfare entry, regardless of whether the individual leaves welfare during that period.

For an individual who leaves welfare for at least 1 quarter and then returns, we must decide how to treat each entry onto welfare. Eliminating subsequent welfare entries after the first observed entry would omit later welfare entries but not earlier ones. We therefore treat each entry onto welfare separately, counting the 16 quarters from that entry even if those same quarters are also included in the period following a prior or subsequent entry.[7] The analysis should thus be properly viewed as identifying earnings outcomes following a particular entry onto welfare.

We control for the extent of welfare experience in the 2 years prior to welfare entry, but we do not control for past participation in welfare-to-work activities or other training. Hence, the estimated impact of welfare-to-work (or work component)[8] participation is an incremental impact, indicating the effect beyond any training received prior to entering welfare. These estimates thus address the appropriate policy question of how the "average" welfare recipient's earnings trajectory is affected by these welfare-to-work program activities.

In examining estimated impacts of the program, we group work components into three categories based on their relative intensity: (1) assessment, (2) job search/readiness training, and (3) intensive training. Earnings impacts are assumed to result from the most intensive service received in the period since entering welfare. We provide an extensive discussion of this categorization in the subsequent section, in part to explicate how our measures of program activities improve on those of previous studies.

Our use of administrative data allows us access to very large samples.

[6] We do not have a full 16 quarters of follow-up data for those entering TANF near the end of our entry window.

[7] We found that, in both Missouri and North Carolina, approximately 1 in 10 quarters in our analysis appears twice, with less than 1% of earnings quarters appearing more than twice.

[8] We use the term "work component" to refer to the particular components or subprograms of welfare-to-work program activities.

In Missouri, we base our analyses on 60,483 unique individuals who meet our sample criteria and who enter welfare a total of 69,551 times during our sample window. In North Carolina, we have 73,837 unique individuals, who enter welfare 82,056 times.[9] In each state, our estimates of program effects are based on more than 800,000 quarters of earnings data for these individuals.

IV. Work Component Activities

The emphasis on moving welfare recipients to work began to take concrete form in the early 1990s with the implementation of the federal JOBS program, which required states to provide explicit services to recipients of Aid to Families with Dependent Children (AFDC). These programs expanded during the decade under federal waivers to states that allowed modification of the AFDC program and then under the federal reforms that replaced AFDC with TANF.[10] North Carolina was an early implementer of these reforms, emphasizing a work-first approach that focused primarily on getting recipients into jobs and secondarily on training to improve skills. North Carolina's TANF program began in January 1997 with a "primary focus" on "job placement assistance."[11] Missouri's approach was less clear. Having emphasized long-term training under JOBS, Missouri's program was modified in the direction of work-first only in the face of federal pressure implicit in the TANF rules. Nonetheless, Missouri's TANF program, which began in December 1996, retained a greater emphasis on long-term training, and by 2000, Missouri had managed another policy turnaround, adopting rules that increased the ability of TANF recipients to engage in long-term training.

Work Component Categories

We have classified the various work component activities into six categories that allow comparability between Missouri and North Carolina. Table 1 provides basic information on the character of these activities. We present statistics on the duration of each activity and the number of hours per week of participation normally scheduled. We have calculated duration as the number of weeks between the date the activity commences and the date when it is completed.

As expected, there are substantial differences between activities in their duration and intensity, as well as differences between states. The first

[9] These counts omit individuals with missing data on variables used in our analysis or with invalid social security numbers. Fewer than 0.5% of data were omitted for these reasons in either state.

[10] See Grogger and Karoly (2005) for a comprehensive summary of state policy changes in the transition from AFDC to TANF.

[11] North Carolina's TANF Web site describes its approach as "grounded in the 'work-first' philosophy." See http://www.joblink.state.nc.us/centers/resources.asp.

Table 1
Work Component Activities Duration and Intensity

	Missouri				North Carolina			
	Duration (Weeks)				Duration (Weeks)			
Activity	25th Percentile	Median Duration	75th Percentile	Median Hours per Week	25th Percentile	Median Duration	75th Percentile	Median Hours per Week
Assessment	1.6	4.9	59.7	20	.0	1.4	4.3	3.0
Job search and job readiness training	1.6	4.4	23.0	25	2.4	4.9	10.3	20.0
Work experience	1.9	6.4	28.1	20	2.4	5.7	12.1	26.0
Basic education	1.6	5.9	30.9	20	4.1	8.4	15.0	20.0
Vocational and technical skills training	2.6	9.4	44.4	25	2.7	6.3	13.0	35.0
Postsecondary education	3.6	19.3	116.4	17	7.4	13.8	24.5	42.5

NOTE.—Statistics are based on all work component activities that begin in 1997 Q2–2000 Q2 in Missouri and 1997 Q2–2001 Q4 in North Carolina for TANF payees who are females, aged at least 18 but less than 65, in the single-parent program, and not in child-only cases.

category, assessment, may include formal paper-and-pencil testing, as well as development of a "self-sufficiency plan," which provides a schedule of activities leading to employment and exit from TANF. In North Carolina, these activities usually take around 3 hours per week and extend for less than 2 weeks. In Missouri, both the reported duration and intensity of assessment activities are greater, but the longer duration is likely at least partly due to systematic errors in data entry.[12]

Job search and job readiness components appear to have similar levels of intensity in both locations, although the upper tail is much higher for Missouri, likely reflecting data errors. The types of activities defined as "work experience" may differ appreciably across programs. Nonetheless, the patterns of participation are similar in the two states, again with the exception of Missouri's longer upper tail.

Basic education includes attendance in public schools up through twelfth grade and English-as-a-second-language instruction, although the largest category by far is adult education and literacy programs, such as those preparing individuals for the high school equivalency diploma. Interestingly, the median number of weeks is slightly greater in North Carolina than in Missouri, although there are more individuals with very long recorded involvement in Missouri.

The typical vocational and technical skills training component lasts about 9 weeks in Missouri but only 6 weeks in North Carolina. In the case of postsecondary education, the median involvement is about 20 weeks in Missouri but only 14 in North Carolina. Yet, the number of hours of involvement per week is much higher in North Carolina, with the median over 40 hours as compared to less than 20 in Missouri, which is very likely due to coding differences.[13]

Despite differences in the duration of training, it is worth noting that the median duration of participation is short in all activities, less than 10 weeks in every category except for postsecondary education. Differences in duration among recipients in a particular type of activity are greater in Missouri than in North Carolina, due largely to the longer upper tail

[12] Although case managers are formally required to specify the date when assessment is completed, in practice they may frequently fail to enter it. In some cases, this may occur when individuals are classified as exempt or are removed from the program for reasons unrelated to program participation. By statute, assessment can take no more than 30 days, in contrast to a median reported assessment time of 4.9 weeks. We were told that apparent deviations from the 30-day limit very likely reflect entry errors. We suspect that end dates for other work component activities may be recorded with error as well.

[13] In Missouri, caseworkers are instructed to include in the scheduled hours 1 hour of study for each class hour, so 17 scheduled hours would indicate 8.5 hours of classes per week, more than half-time in most colleges. We suspect that the 42.5 hours per week scheduled in North Carolina reflects a more liberal coding for study time.

in the Missouri distribution. In the analysis that follows, we will focus on the effects as measured from the quarter of initial participation in an activity. It must be understood that earnings in the quarter in which individuals participate, and possibly subsequent quarters, will be depressed if there is a period of extended participation and the program precludes or substitutes for immediate employment.

We use this approach, rather than considering the time after completion, for several reasons. First, we are unsure about the validity of program end dates, especially in the Missouri data (see n. 12). Second, and perhaps more important, we are concerned that, even if the exit date were properly coded for an individual spending an extended period in a component, the end date might be endogenous with employment success. It would not be surprising if those registered for an extended period would be discontinued when they obtained employment (see, e.g., Courty and Marschke's 2004 analysis of JTPA program exits). Third, we believe that even where measures of time in a program are not problematic, program impact is more meaningfully measured if it applies to periods during participation in the program.

This approach treats time spent in a program as comparable with other factors that may delay movement into productive employment. Implicitly we are assuming that a program lasting 6 months that immediately places individuals in appropriate employment is not necessarily better than a 3-month program that requires individuals to spend 3 additional months obtaining suitable employment. In the discussion that follows, when we refer to the time of program participation or receipt of services, we are referring to the quarter when participation in the activity first began. Although for a large share of participants measured program completion occurs within that same quarter, for some individuals it will extend to subsequent quarters.

In order to avoid problems associated with small numbers of observations, we group together activities in the bottom four categories in table 1 as "intensive training" or "intensive services." Although there is an appreciable variation among them, in both states median duration is longer for each of these categories than for assessment or job search/readiness training. Our use of three categories of participation—assessment, job search/readiness training, and intensive training—also allows us to easily compare the effect of work-first activities, such as job search, with more intensive activities, such as vocational education, that are designed to enhance participants' human capital.

Table 2 provides information on the ordering of component participation for individuals participating in at least one component in the 16 quarters since entering welfare. For each case, we consider only the first occurrence of an activity, and we count activities occurring within the same quarter as ties. For each row, the base is the number of welfare

Table 2
Component Order (%)

	1	Tie (1)	2	Tie (2)	3	Total
Missouri:						
Assessment	51.7	34.6	9.8	2.1	1.8	100.0
Job search/readiness	46.8	25.6	15.2	2.2	10.2	100.0
Intensive training	30.2	39.9	20.5	.0	9.5	100.0
North Carolina:						
Assessment	52.6	45.6	1.0	.3	.5	100.0
Job search/readiness	7.3	56.8	17.0	6.5	12.4	100.0
Intensive training	5.4	52.3	17.1	7.5	17.7	100.0

NOTE.—Each row is based on the number of individuals entering welfare who participated in the specified service at some time in the 16 quarters after entry. Ordering is determined by quarter of participation, with ties identifying participation in the same quarter (see text). Individual rows may not sum to 100 due to rounding error.

entries in which the listed activity occurred in our 16-quarter window. For example, we see that for cases where assessment occurred in Missouri, it occurred (for the first time) in a quarter prior to any other activity in 52% of the cases.[14] In an additional 35% of cases, assessment occurred in the same quarter as the first occurrence of another activity. This means that of those who receive assessment, 86% in Missouri receive it in the first quarter that they receive any component. In North Carolina, the comparable figure is 98%.

The two states differ more dramatically, however, in the likelihood that an individual who participates in job search/readiness or intensive training does so in a quarter prior to any other activity. In Missouri, nearly half of those who participate in job search/readiness do so in a quarter prior to receiving any other service, and nearly a third of those in intensive training participate in a quarter prior to receiving any other training. In contrast, fewer than 10% of those in North Carolina participate in these more intensive services without also participating in another component—usually assessment.

Since a large share of participants enter more than one type of component, we must decide how to gauge impacts in such cases. A simple additive model assumes that a component contributes to outcomes without regard for whether it is combined with other components. Such an approach would require that we decide how individuals who participate in more than one component within our categories are treated. In keeping with our focus on the impact of component intensity, we identify the type of training by the highest intensity component that the individual participated in since coming onto welfare. In particular, a quarter is coded as "assessment only" if the individual received assessment services at some point since coming onto welfare but has not received any other work

[14] This includes those cases where it was the only activity.

Table 3
Outcome Earnings Coding Example

Quarter	Service Received	Service Code	Observed Earnings
0	No service	$s_{i0} = 0$	$Y_{i0} = Y^0_{i0}$
1	No service	$s_{i1} = 0$	$Y_{i1} = Y^0_{i1}$
2	Assessment	$s_{i2} = 1$	$Y_{i2} = Y^1_{i20}$
3	No service	$s_{i3} = 1$	$Y_{i3} = Y^1_{i21}$
4	No service	$s_{i4} = 1$	$Y_{i4} = Y^1_{i22}$
5	Job search/readiness	$s_{i5} = 2$	$Y_{i5} = Y^2_{i50}$
6	No service	$s_{i6} = 2$	$Y_{i6} = Y^2_{i51}$
7	Intensive services	$s_{i7} = 3$	$Y_{i7} = Y^3_{i70}$
8	No service	$s_{i8} = 3$	$Y_{i8} = Y^3_{i71}$
9	Job search/readiness*	$s_{i9} = 3$	$Y_{i9} = Y^3_{i72}$
10	No service	$s_{i10} = 3$	$Y_{i10} = Y^3_{i73}$
⋮	⋮	⋮	⋮
15	No service	$s_{i15} = 3$	$Y_{i15} = Y^3_{i78}$

NOTE.—$Y_{it'}$ is observed earnings in quarter t' for individual i; $Y^0_{it'}$ is earnings in quarter t' conditional on no participation in any service since entering welfare; Y^s_{itd} is earnings in quarter $t' = t + d$, conditional on participating in service s ($s = 1$, 2, or 3) in quarter t, d quarters earlier; $s_{it'}$ is the highest-level service received since entering welfare to time t'. All quarters are specified relative to the quarter when entering welfare (quarter 0).
* This service has no impact on the coding.

component services. A quarter is coded as "job search/readiness" in quarters following receipt of such services, as long as the individual has not received any "intensive" services to that point. Finally, a quarter is coded as "intensive services" following the receipt of intensive services, without regard for whether any other services were previously received. Hence, the intensive service category includes any effects of other services received by such individuals, and impact estimates must be interpreted accordingly. Table 3 illustrates the way in which services are coded.

Timing of Program Participation

The TANF recipients are most likely to participate in work components in the first year after entering the program. Table 4 indicates job component participation by quarter since entering welfare for the sample of cases entering welfare in 1997 Q2–Q4. In keeping with our hierarchical coding, we identify only the highest-order activity up to the quarter indicated. Column 6 shows that in Missouri, 10% of the sample participates in some component in the entry quarter; for North Carolina, this number is 9%. In both states, the proportion who had participated in at least one component increases to over 20% in the quarter following entry and exceeds half by the end of our 16-quarter period. Of the individuals who had participated in at least one work component in the 4 years after entering TANF, about 80% had participated in the first 2 years in both states.

The chance of participation declines over time in part because, after several quarters, a substantial portion of recipients have left welfare.

Table 4
Component Cumulative Participation by Quarter after TANF Entry (%)

Quarter after TANF Entry	No Component and Receiving TANF (1)	No Component and Exit from TANF (2)	Assessment Only (3)	Job Search/Readiness and No Intensive Training (4)	Intensive Training (5)	Total Participating in Any Activity* (6)
Missouri:						
0	90.2	.0	5.1	1.8	3.0	9.8
1	64.8	14.8	8.9	4.0	7.6	20.4
2	44.8	28.8	11.3	5.4	9.8	26.4
3	32.8	36.4	12.6	6.6	11.6	30.8
5	19.4	43.5	13.5	9.3	14.3	37.1
7	11.6	46.3	13.7	11.7	16.8	42.2
11	5.9	45.9	14.2	13.9	20.1	48.2
15	3.3	44.7	14.2	14.6	23.2	52.0
North Carolina:						
0	90.9	.0	6.2	1.7	1.2	9.1
1	65.9	11.1	13.1	5.0	4.9	23.0
2	40.2	31.6	14.6	6.4	7.2	28.2
3	26.7	41.2	15.8	7.4	8.9	32.1
5	13.7	47.6	17.4	9.3	12.0	38.7
7	7.2	49.6	17.8	10.4	14.9	43.2
11	2.6	49.4	18.1	11.3	18.6	48.0
15	1.6	48.0	17.9	11.7	20.8	50.4

NOTE.—Columns 1 and 2 identify individuals who have not participated in any work component since entering welfare; the former includes only individuals receiving welfare during the specified quarter; the latter includes individuals who are not receiving welfare. A small portion of recipients included in col. 1 will have exited following the initial welfare entry and then reentered welfare. Individuals are followed for 16 quarters after entry without regard to whether they exited welfare during the period. Columns 3–5 may not sum to col. 6 due to rounding errors.

* Sum of cols. 3–5.

Those who have not participated in any component are distinguished in columns 1 and 2 in the table by whether they were receiving TANF in the specified quarter or had exited welfare. Comparing the columns, we see that by quarter 3 most who had not participated in any component had left welfare, and by quarter 11, about 95% of those who had not participated had left welfare. These patterns in part reflect the strict participation requirements for welfare recipients following the 1996 welfare reforms. In turn, it is clear that the welfare exit rate has an important impact on the pattern of overall participation in work component activities, since almost everyone who continues to receive welfare participates in a work component within several years.

In both states, of those participating in a component in their first quarter, more than half participate in assessment only, while about a quarter in both states are listed as participants in job search/readiness (they may or may not have participated in assessment). More than twice as many participants in Missouri (as in North Carolina)—nearly a third of those participating in some component—participate in an intensive activity in the first quarter. As we look at later quarters, we see that there is greater involvement in intensive activities in Missouri but that by the end of the 16 quarters most of the difference has disappeared. It is useful to note that in both Missouri and North Carolina the number of recipients coded as having participated in assessment remains steady after about the eighth quarter. This implies that, although new participants may be assessed each quarter, an equal number of those who were assessed in prior quarters are receiving other services.[15]

Program Participation and Earnings Patterns

It is well known that individuals who enter training and related programs for the disadvantaged often experience earnings declines associated with initial program participation, reflecting program requirements as well as self-selection. Individuals in our sample enter into welfare and then participate in work component activities, and thus we might anticipate that this Ashenfelter's dip would be observed for both. Figures 1 and 2

[15] In an earlier version of the article, we also examined how participation had changed for those entering welfare during the period of our study (Dyke et al. 2005). We saw in both states that the chance of participation in at least one component increased from about 40% to over 50%. The two more intensive activities showed substantial increases, both increasing by more than 50%. The assessment category did not increase over time. This is a result of the fact that, although more individuals are, in fact, receiving assessment services, the growth in other services hides this. Our analysis of work component participation suggests that the differences in welfare-to-work strategies adopted by North Carolina and Missouri are primarily rhetorical and that patterns of participation in the different activities are not only similar but changing in similar ways over time.

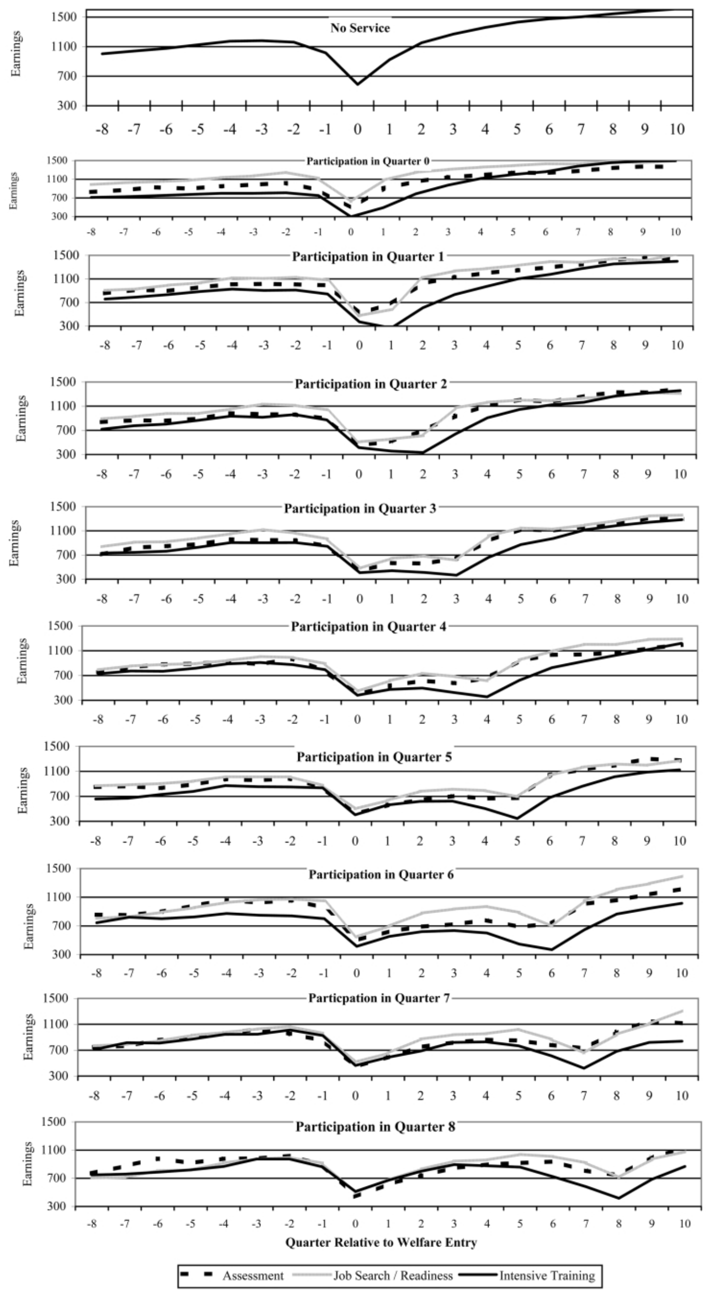

Fig. 1.—Patterns of average quarterly earnings for participants in work components by quarters of participation in Missouri.

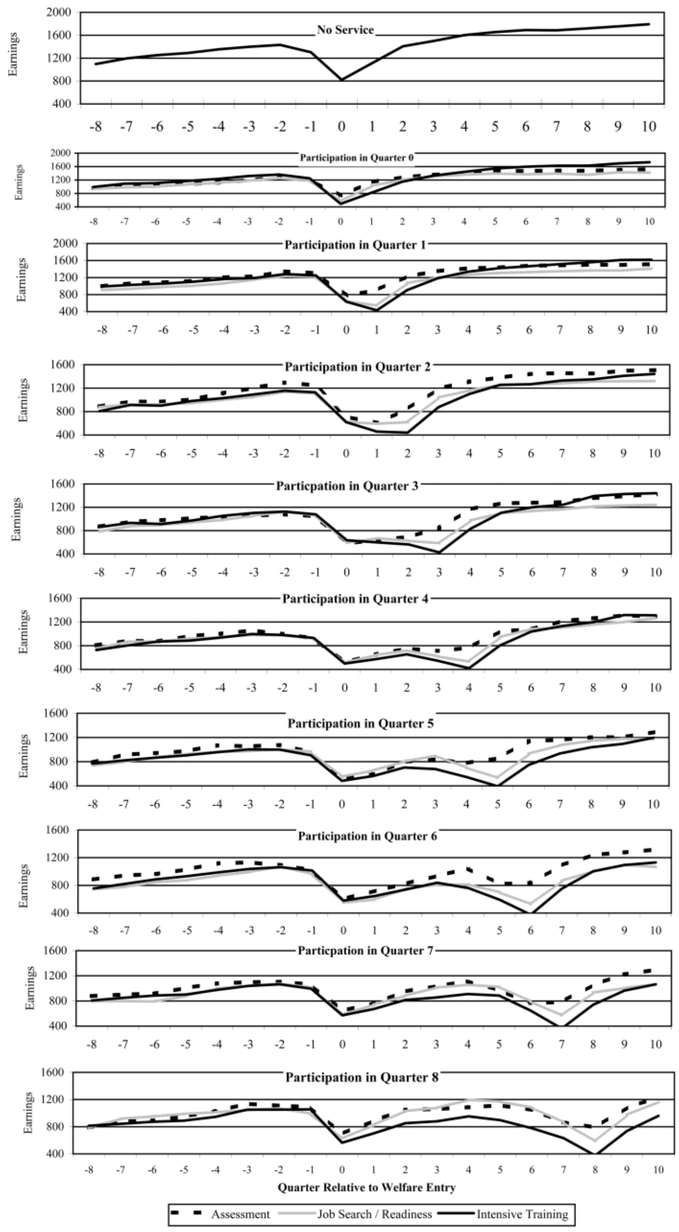

FIG. 2.—Patterns of average quarterly earnings for participants in work components by quarters of participation in North Carolina.

show average earnings for Missouri and North Carolina, respectively, over 8 quarters prior to welfare entry and the subsequent 10 quarters.

The top graph shows earnings for individuals who do not participate in any work component over the 16 quarters beginning with the quarter of welfare entry. Looking at the second quarter prior to program entry, we see that earnings are about \$1,200 in Missouri and \$1,400 in North Carolina, but in both states there is a decline of approximately \$600 in the quarter of entry onto welfare. Over the subsequent 10 quarters, we see that earnings increase, ultimately exceeding \$1,600 in both states.

For the lower graphs in the figures, average earnings are identified for individuals who participate in at least one work component activity, separated by the starting quarter of that activity. In both states, we see two earnings dips, one for welfare entry and the other for participation. Where work component participation occurs within 3 quarters of welfare entry, earnings decline from welfare entry to quarter of participation. For later participation, there is an increase in earnings followed by a decline in the quarter of participation. Especially for those receiving services several quarters after welfare entry, the decline in earnings in the quarter of participation is greatest for those participating in intensive training. Generally, the decline for assessment is least severe. Perhaps most remarkable is the similarity in patterns between the states. These patterns underscore the importance of taking into account individuals' selection into welfare as well as possible selection into work components.

V. Models of Impact and Estimation Strategy

Entry onto welfare is selective not only of particular kinds of individuals but also along a temporal dimension for these individuals. For individuals who are normally employed, entry onto welfare will occur in a period where there is an unexpected negative shock. Insofar as such circumstances are not perfectly correlated over time, subsequent earnings would be expected to increase even in the absence of government intervention. Therefore, we control for the time since entry onto welfare. Our approach accounts for regression to the mean and for the impacts of the TANF program that may not be captured in work component participation.

The model underlying our estimation strategy can be written as

$$Y^s_{itd} = Y^0_{it'} + \delta^s_{td}(X_{it'}) + \varepsilon^s_{itd}, \tag{1}$$

where Y^s_{itd} is earnings for individual i, d quarters after receiving service s in quarter t, where no higher-level service was received up through $t' = t + d$ (see table 3 for an example of how earnings are coded). All times are expressed relative to the time coming onto welfare, so $t = 0$ implies that the service is received in the quarter in which the individual entered welfare, and $t = 1$ implies the service is received in the first quarter following entry, and so on. Individuals' characteristics and other factors (e.g., the county unemployment rate) for the outcome quarter t' are captured in $X_{it'}$, $\delta^s_{td}(X_{it'})$ is the expected impact of the service in quarter $t' = t + d$ expressed as a function of individual characteristics and calendar quarter (captured in $X_{it'}$), and $Y^0_{it'}$ is expected earnings in quarter $t' = t + d$ in the absence of participating in any work component since entering welfare. The error term is given by ε^s_{itd}, and we assume that $E\left(\varepsilon^s_{itd}|X_{it'}\right) = 0$.

The structure of effects on earnings in this model is quite general, in that characteristics may influence earnings in a way that can vary idiosyncratically across service and outcome quarter. For individual i, the impact of program participation in t on earnings in quarter $t' = t + d$ is given as

$$Y^s_{itd} - Y^0_{it'} = \delta^s_{td}(X_{it'}) + \varepsilon^s_{itd}. \tag{2}$$

We are primarily interested in estimates of the average impact of participating in service s in quarter t on earnings d quarters later for those who actually receive services, that is, the effect of the treatment on treated individuals. This can be written as $\delta^s_{td} = E[\delta^s_{td}(X)]$, where the expectation is across the characteristics of participants who receive services in t and who have received no higher-level services in the following d quarters. We also consider the expected impact of service on earnings d quarters later, averaging across quarters in which the service was received. We write this as $\delta^s_d = E[\delta^s_{td}(X)]$. This average is across participants receiving services in all quarters who had not received higher-level services in the following d quarters.

For any one individual, there are a large number of potential wages conditioned on the receipt and timing of services, but we only observe those wages associated with services that were actually received, that is, we observe

$$Y_{it'} = Y^0_{it'} + \sum\nolimits_{s\in\{1,2,3\};t\in\{t|0\leq t\leq t'\}} D(s_{it'} = s \text{ and } t_{it'} = t)(Y^s_{itd} - Y^0_{it}),$$

where $s_{it'}$ ($s_{it'} = 0$, 1, 2, or 3) is the highest-order service actually received by individual i from the quarter of welfare entry through quarter t', $t_{it'}$ is the quarter in which that service was first received, $d = t' - t$, and D is an indicator dummy taking on a value of one to identify the particular treatment received and its timing. If the individual has not received any service up through t', then $s_{it'} = 0$ and the summation drops out.[16]

We estimate our model using propensity score matching. Matching methods are more general than linear models in that they provide estimates of program effects that relax several assumptions of the linear model.

[16] When $s_{it'} = 0$, we arbitrarily set $t_{it'} = 0$.

First, these models recognize that program effects may differ across individuals, explicitly producing estimates that are averages in the specified population. In addition, linear regression models may perform poorly when participants and comparison groups have significantly different values on control variables. In the extreme, some treated cases may not have comparable matches in the comparison sample. A matching approach allows us to identify such failures of common support.

The critical assumption necessary in order to apply matching methods to obtain estimates of the treatment effect on the treated individuals is that participation in a work component activity and its timing are conditionally independent of the earnings that would be received in the absence of participation. Since the earnings an individual would receive during this quarter in the absence of any participation are $Y^0_{it'}(t' = t + d)$, the independence condition can be written as

$$Y^0_{it'} \perp (s_{it'}, t_{it'}) | X_{it'}, t', \tag{3}$$

where $s_{it'}$ and $t_{it'}$ identify the services and timing of participation for i, as defined above. Given this assumption, for a given case i, it is possible to obtain an unbiased estimate of $Y^0_{it'}$ for an individual i who received service $s_{it'} > 0$ by using the earnings of individuals who received no services up through quarter t' but who have the same values of $X_{it'}$. We write the estimate of program impact for individual i as

$$\hat{\delta}^s_{itd} = Y^s_{itd} - Y^0_{t'}(X_{it'}), \tag{4}$$

where $Y^0_{t'}(X_{it'})$ is estimated earnings in outcome quarter $t' = t + d$ based on comparison cases receiving no services through quarter t' that are matched with the participant case i.

We use radius matching based on propensity scores, incorporating a bias adjustment procedure. Given that the timing of participation in a component may not be random, we use a propensity score matching approach structured to reflect the dynamics of the participation decision. Our method is similar to that used by Sianesi (2004) and Fitzenberg, Völter, and Osikominu (2006) in similar settings. The following section describes our methods in greater detail.

We obtain average impact estimates as the mean impact across all participants

$$\hat{\delta}^s_{td} = (1/N_{td}) \sum_{i \in T_{td}} \hat{\delta}^s_{itd},$$

where T_{td} is the set of individuals receiving service in quarter t who have not received any higher-level service d quarters later, and N_{td}, the number of such individuals. We then have

$$\hat{\delta}^s_d = (1/N_d)\sum_t N_{td}\hat{\delta}^s_{td},$$

where $N_d = \sum_t N_{td}$ is the sum of all participants with observed earnings d quarters after receipt of service s.[17]

Controlling for Program Selection

Perhaps the most important challenge to the assumption of conditional independence specified above is that participation in work component programs is, by institutional design, nonrandom, and the factors that enter into this decision are not directly observable. The possibility that selection may play a role in the earnings outcome we observe is underscored by the patterns we observe in figures 1 and 2, which show that participants experience an earnings dip in the quarters prior to participation in a work component.

Whether a TANF recipient is required to participate in a work component activity, and the activity that is recommended, depends on the circumstances of the recipient. Program rules exempt certain recipients from participation, such as individuals with very young children. Other exemptions are based on the judgment of the caseworker, as when an individual is viewed as facing personal obstacles that make it too difficult to engage in training or employment.

There is also an element of personal choice. In Missouri, individuals who fail to participate in required programs face sanctions that reduce their payments (generally by about 25%) but are permitted to continue receiving these reduced TANF benefits until the 5 years of eligibility is exhausted. However, in North Carolina, those who fail to cooperate can have the full value of their benefit withheld. Equally important, in both Missouri and North Carolina, individuals who work a minimum number of hours are exempted from participation in work component activities. As a result, those who participate may be individuals whose labor market

[17] In a previous version of the article (Dyke et al. 2005), we employed a simplified version of eq. (1), $Y^s_{itd} = X_{it'}\beta + \gamma_{t'} + \delta^s_d + \varepsilon_{itd}$. We estimated δ^s_d both with ordinary least squares (OLS) and an instrumental variables estimation approach, using variation over time in the level of participation in training programs across counties as the identifying variable. Instrumental variable estimates were generally implausible, suggesting that the identifying variable was correlated with unmeasured labor market differences. Simple OLS estimates of δ^s_d were generally reasonable, but in some cases they were not robust when the model was fitted on alternative subsamples, suggesting that the specification was too restrictive. Although we could have developed more general linear models, incorporating interaction effects, a matching strategy both implies a substantially more general model and explicitly takes account of differences in program impact across individuals.

opportunities are particularly limited or who are facing an extended streak of bad luck.

We consider a more plausible alternative model that acknowledges the possibility of selection on unmeasured factors. This model makes two additional assumptions, first, that selection into service is the same for the three classes of service and, second, that assessment provides minimal benefits to those who receive only assessment:

$$Y^0_{it'} \perp s_{it'} | X_{it'}, t_{it'}, t', \text{ for } s_{it'} > 0;$$
$$\delta^1_{td}(X_{it'}) = 0. \tag{5}$$

In essence, the first expression specifies that individuals participating in a given quarter are selected in the same way regardless of the service they receive. The underlying assumption is closely related to that made by Hotz et al. (2006, in this issue) to identify the relative importance of alternative services. By conditioning on the quarter of service, we allow for the possibility that selection of individuals may differ by quarter of service (relative to the quarter of welfare entry) and that it may differ dramatically for those receiving no services.

The assumption that participation in assessment alone provides no benefits in later periods is reasonable on its face because of the limited time individuals spend in assessment. Given this assumption, an unbiased estimate of $Y^0_{it'}$ for an individual i who received service $s_{it'} > 1$ in quarter $t_{it'}$ and received no higher-level service through t' can be obtained from the earnings of individuals who received assessment in quarter $t_{it'}$ and no higher-level services up through quarter t' but who have the same values of $X_{it'}$. For individual i who received services s ($s = 2$ or 3) in quarter t, the estimate of program impact on earnings d quarters later can be written as

$$\hat{\delta}^s_{itd} = Y^s_{itd} - Y^1_{td}(X_{it'}), \text{ for } s > 1, \tag{6}$$

where $Y^1_{td}(X_{it'})$ is the estimated earnings in quarter $t' = t + d$, based on matched cases that participate in assessment in quarter t.

Accounting for Individual Fixed Effects

Individual fixed-effects estimators provide an alternative approach to controlling for differences across individuals who participate in different kinds of services. This approach, in essence, produces estimates of the impact of work component participation by comparing a recipient's experience prior to component participation with her subsequent experience. Smith and Todd (2005) spell out the basic approach, which they describe as "difference-in-difference" matching. For treated cases, the dependent variable is the difference between earnings in a period following participation and earnings prior to program participation, and for comparison cases the earnings difference is calculated over the same periods. Even if individuals who obtain services differ in important ways from those in the comparison group, so long as such differences are stable over time in their impact on earnings, this specification can eliminate bias resulting from differences between participants and others. The approach can be illustrated by returning to the initial specification in equation (1) and subtracting earnings prior to service receipt from both sides of the equation, obtaining

$$\Delta Y^s_{itd} = \Delta Y^0_{itd} + \delta^s_{td}(X_{it'}) + \varepsilon^s_{itd},$$

where ΔY^s_{itd} and ΔY^0_{itd} are differenced earnings measures for individual i, and where $t' = t + d$. We will take earlier earnings to be average earnings for the quarters after welfare entry and prior to the service quarter t, that is,

$$\Delta Y^s_{itd} = Y^s_{itd} - (1/t)\sum_{t''=0}^{t-1} Y_{it''}, \text{ and}$$

$$\Delta Y^0_{itd} = Y^0_{it'} - (1/t)\sum_{t''=0}^{t-1} Y_{it''}.$$

This approach removes the impact of any fixed individual effect that influences earnings both before and after program participation. If the independence assumption (eq. [3]) is violated because fixed effects differ for those receiving services and the comparison group, the assumption may still hold for the differenced measures; that is, it may be the case that

$$\Delta Y^0_{it'} \perp (s_{it'}, t_{it'}) | X_{it'}, t'.$$

In this case, the differenced earnings for those not receiving treatment can be used to estimate what differenced earnings would have been for those who received some service in the absence of services, that is, in the counterfactual state. The estimated program impact for case i is then written as

$$\hat{\delta}^s_{itd} = \Delta Y^s_{itd} - \Delta Y^0_{td}(X_{it'}), \tag{7}$$

where $\Delta Y^0_{td}(X_{it'})$ is the mean differenced earnings for cases not receiving services up to t' and matched to i on the basis of characteristics $X_{it'}$.

A similar argument may apply in the case where those receiving assessment form the comparison group. If equation (5) is violated because individuals who undertake assessment differ from those receiving more intensive services due to differences in expected fixed effects, then it may be that the condition put in terms of differences may still hold,

$$\Delta Y^0_{it_{it'}d} \perp s_{it'} | X_{it'}, t_{it'}, t', \text{ for } s_{it'} > 0.$$

This expression states that the differenced measure of the earnings individual i would receive in the absence of services is unrelated to which services i actually receives, conditioned on i receiving some services ($s > 0$). Combined with the assumption that $\delta^1_{td}(X_{it'}) = 0$, the estimate of the program $s > 1$ effect for case i is given as

$$\hat{\delta}^s_{itd} = \Delta Y^s_{itd} - \Delta Y^1_{td}(X_{it'}), \text{ for } s > 1, \tag{8}$$

where $\Delta Y^1_{td}(X_{it'})$ is the mean differenced earnings for cases that receive assessment in t, receive no other services in the next d quarters, and are matched to i on the basis of characteristics $X_{it'}$.

Specification Tests

Each of the above approaches makes certain assumptions in order to obtain estimates of the effects of service receipt. One natural specification test is to use the same methods to predict the effects on earnings that precede participation in the program (e.g., Mueser, Troske, and Gorislavsky 2005). If we obtain nonzero impact estimates for these periods, this suggests that our methods are not properly accounting for unmeasured differences between those cases receiving services and those used in comparisons. In particular, we estimate the effects of service on earnings in the quarters after entering welfare but prior to receiving a service. Consider the estimate of the impact of a service s in quarter t on earnings in quarter $t' = t + d < t$, that is, where $t > |d| > 0$,

$$\hat{\delta}^s_{itd} = Y^s_{itd} - Y^0_{td}(X_{it'}), \tag{9}$$

where Y^s_{itd} is the earnings in quarter t' for an individual receiving services in quarter t, and $Y^0_{td}(X_{it'})$ is the earnings in t' based on cases matched to i that receive no services up through at least t. Recall that the variables included in $X_{it'}$ are the characteristics of the individual, earnings prior to entering welfare, and characteristics of the economy during the outcome quarter, but earnings following welfare entry and prior to t are not included. Hence, our methods do not impose any mechanical requirement that matched cases have similar values on earnings in the quarters between welfare entry and service receipt. If estimated “effects” on prior earnings are close to zero, this suggests that important differences may be largely controlled by our methods, supporting the validity of our impact estimates.[18]

The same approach can be used to test the fixed-effects specification (Heckman and Hotz 1989). Equations (7) and (8) provide estimates of program effects for quarters following a service in quarter t. However, for such a case, if one specifies an outcome quarter $t' < t$ and a "pseudo" service quarter $t^* \leq t$, an estimate of the "impact" of a service on prior quarters can be obtained. In this case, the outcome measure is earnings in quarter t' minus average earnings in quarters 0 through $t^* - 1$. Impact estimates are based on comparing outcome measures for such treated cases (in this example, cases receiving treatment in quarter t) with identically calculated measures for matched comparison cases receiving either no treatment (eq. [7]) or assessment (eq. [8]) in quarter t.

VI. Estimation Details and Specification Test Results

As indicated above, our impact estimates are based on propensity score matching. We identify a treatment as a service s received in quarter t influencing earnings in quarter t', which is $d = t' - t$ quarters following treatment. We specify $t = 0$ as the quarter of entry onto welfare, and we limit consideration to treatments involving service quarters up through $t = 10$, since a relatively small proportion of services are received after the tenth quarter. We consider $t' = 0$–15, so we examine 16 outcome quarters beginning with the quarter of entry to welfare.[19]

For a given treated subsample (defined by t, d, t', and s), we identify a set of comparison individuals who contribute quarters such that time since entry is identical to the treated subsample. The estimates based on equations (4) or (7) use earnings in quarter t' for individuals who participated in no work component activity up to that point. The estimates based on equations (6) or (8) use a comparison sample of individuals who participated in assessment in quarter t and are observed to have received no higher-level services through quarter t'. This latter comparison sample is the same as the treatment subsample (t, d, t', and 1).

In undertaking propensity score matching for a given treatment subsample and its comparison group, we fit a logit using $X_{it'}$ to predict whether a case is a treated or a comparison case and then use estimated coefficients to construct a predicted probability or "propensity score" for

[18] A similar specification test is easily constructed based on the comparison of services with assessment. Here $\hat{\delta}^s_{itd} = Y^s_{itd} - Y^1_{td}(X_{it'})$, where again we take outcome quarter $t' = t + d < t$, and $Y^1_{td}(X_{it'})$ is the earnings in quarter t' for individuals receiving assessment in quarter t who are matched to individual i.

[19] Recall services are coded as $s = 0$ for no service, $s = 1$ for assessment, $s = 2$ for job search/readiness training, and $s = 3$ for intensive training activities, where codes identify the highest-level service received to that point after entering welfare (see table 3).

each case. We include a fairly extensive set of variables in this model, identifying both individual characteristics and prior labor market experience.[20] We also enter dummies identifying the calendar quarter of the outcome, the county unemployment rate during that quarter, and dummies identifying 17 county groups in Missouri and 24 county groups in North Carolina. Subject to parametric assumptions and small numbers limitations, a treated and comparison case with the same propensity score will have the same distribution of values on X_{it} (Rosenbaum and Rubin 1983); that is, the estimated propensity score is a balancing score.

We employ radius matching in order to identify comparison cases that correspond with our treated cases. For each treated case within the subgroup, we designate as “matches” all comparison cases that have propensity score values that are within 0.005 of the treated case.[21] This method not only allows for more than one comparison case to be matched with a treated case but, because the matching search in the comparison sample is done with replacement, it also allows a given comparison case to be matched with more than one treated case. The mean outcome for cases matched with a given treated case is an estimate of the outcome that would occur for the treated case in the absence of the service.[22]

Since the propensity score is based on fitting a parametric structure, it is necessary to test to assure that the estimated propensity score is successful in balancing values of matched treatment and comparison cases (see Smith and Todd 2005). Following the matching, within a given treatment subsample, we compared the means of each variable for the treated cases and the weighted comparison cases. Since there are over 40 variables and up to 110 treatment subsamples for each comparison, this implies as

[20] Personal and family characteristics include age and age squared, nonwhite, a dummy identifying education less than high school, number of children, and age of the youngest child. Based on the 2 years prior to entering welfare, we include variables identifying the proportion of quarters worked, a dummy for no work, a dummy for working all 8 quarters, total earnings in the 4 quarters immediately prior to entering welfare, earnings in the fifth through eighth quarters prior to entering welfare, and the proportion of the 8 quarters prior to the observed entry in which welfare was received.

[21] We tested alternative radius values using least squares leave-one-out cross-validation methods (Black and Smith 2004) and found that a radius of 0.005 performed at least as well as larger values.

[22] A recent paper by Abadie and Imbens (2006) argues that, in general, matching estimates include a bias term that disappears relatively slowly as sample size increases. They suggest that “bias adjustment” based on a linear model can improve the performance of estimates. Our estimates of impact include a bias adjustment term to correct for any differences between treated and matched comparison cases. The coefficient used for adjustment is based on a regression within the weighted comparison cases, estimated separately for each treatment subsample. If matching is successful, the adjustment will be small. The bias adjustments had little substantive impact on our results.

many as 5,000 comparisons for each comparison made. We found that our initial logit specification, in which the log odds of the probability was assumed to be a linear, additive function of our variables, was successful at balancing the variables. Overall, we found that far fewer than 5% of the differences were statistically significant at the 0.05 level, implying that matching was more successful than chance at balancing the specified variables.[23]

Frequently, matching estimates are limited by a failure of overlap in the distributions of variables for treated and comparison cases. Although our use of a relatively small matching radius of 0.005 helps to assure comparability for matched cases, it increases the possibility that not all treated cases will be matched. In fact, because of the large number of comparison cases, over 97% of treated cases were matched for each set of outcome estimates reported below.[24]

Conventionally, standard errors of propensity score matching estimates are obtained using bootstrap methods, since there is no analytical formula that accounts fully for the influence of sampling error on propensity score matching estimates. However, where samples sizes are very large, as ours are, the bootstrap is impractical (e.g., Lechner 2001). We therefore present analytical standard errors based on the simplifying assumption that the estimate $\hat{\delta}^{s}_{itd}$ is an independent draw across all treated cases i.

To determine how these analytical standard errors compare with those obtained from a bootstrap, we calculated bootstrap standard errors using 100 replications in analyzing the effect of assessment on earnings 3 quarters after receipt of services. For estimates based on the 10 subsamples (identifying a quarter of service receipt relative to welfare entry), the analytical standard errors were smaller than the bootstrap standard errors, but the difference was less than 10% in all cases but one. For the overall estimate, the analytical standard error was 15% smaller than the bootstrap standard error. This limited comparison suggests that our use of analytical standard errors is unlikely to be seriously misleading.[25]

[23] It is important to recognize that this comparison is not a statistical test. The implication of our results is that differences on measured variables after matching are small relative to sampling error.

[24] In some of the specification tests, the proportion matched was somewhat lower; in one case, as low as 50%. This resulted from our small radius size in conjunction with a small comparison group. Omitted cases appear to be largely random. Note that failure of overlap does not invalidate the specification test.

[25] We expect the true standard error to exceed the analytical standard error for two reasons. First, the analytical standard error cannot account for error in estimation of the propensity score, which may increase the true standard error. Second, our matching methods allow a comparison quarter to be matched with more than one treated case and also allow a given individual to contribute to the comparison sample in more than one subsample. Our bootstrapping methods accounted for both sources of error, because each repetition selects a sample (with

Specification Test Results

Following the structure described above (eq. [9] and n. 18), we have performed specification tests that predict the levels of earnings in the quarters prior to participation in a work component activity but following entry onto welfare. If the methods we use are valid, we should not find significant impacts. Results of these specification tests are reported in table 5. Given our very large sample sizes, these tests were undertaken on a random 25% sample of individuals in the Missouri data and on a 20% sample in the North Carolina data.

Impact estimates for the comparison of participants with a comparison group that received no service are reported in columns 1–3, for Missouri, and columns 6–8, for North Carolina. In most cases, estimated impacts are negative, substantial, and statistically significant. For example, in Missouri, individuals receiving assessment have earnings in the 6 quarters prior to receiving those services that are between \$291 and \$357 below earnings for individuals not receiving services in a given quarter, and estimates for earlier quarters imply a deficit that is generally in excess of \$100, although large standard errors render many of them statistically insignificant. Negative effects are also large for those participating in job search/readiness or intensive activities. Comparable estimates for North Carolina are also substantial and negative for all three categories of service.[26] Clearly, these estimates show that individuals receiving services are systematically different from those who do not receive services. It is therefore unlikely that estimates of program impact on subsequent earnings based on comparisons with nonparticipants will identify true program effects.

Columns 4–5 and 9–10 in table 5 also provide estimates of impact on prior earnings, but these are based on a model that compares those participating in job search/readiness training and those participating in intensive training with those who are only assessed. Looking first at Missouri, for job search/readiness, the estimated impacts of participation on the 6 prior quarters are positive and often statistically significant with quite a bit of variation from quarter to quarter. Estimates for intensive training are usually negative, with the estimate for the prior quarter at −\$115. Estimates for more than 5 quarters prior to participation are generally smaller and not statistically significant. These results suggest that, in the case of Missouri, the model that uses assessment as the com-

replacement) from the sample of individuals and, in each repetition, reestimates the propensity score.

[26] It should be stressed that all the estimates reported in this table identify only quarters after entry onto welfare, because employment prior to labor market entry is controlled in these analyses.

Table 5
Matching Estimates of "Effects" of Services on Earnings in Prior Quarters

	Missouri					North Carolina				
Quarter Relative to Service	Assessment vs. No Service (1)	Job Search/ Readiness vs. No Service (2)	Intensive vs. No Service (3)	Job Search/ Readiness vs. Assessment (4)	Intensive vs. Assessment (5)	Assessment vs. No Service (6)	Job Search/ Readiness vs. No Service (7)	Intensive vs. No Service (8)	Job Search/ Readiness vs. Assessment (9)	Intensive vs. Assessment (10)
−1	−353	−337	−497	103	−115	−322	−447	−610	−117	−184
	(21)	(22)	(17)	(30)	(20)	(21)	(23)	(19)	(34)	(24)
−2	−345	−278	−432	94	−53	−315	−400	−501	−37	−153
	(27)	(26)	(22)	(35)	(30)	(29)	(33)	(24)	(58)	(39)
−3	−357	−250	−375	128	−68	−327	−322	−461	161	−66
	(29)	(29)	(26)	(42)	(35)	(32)	(35)	(27)	(163)	(46)
−4	−302	−192	−312	112	−140	−263	−275	−394	−35	−209
	(36)	(30)	(30)	(59)	(48)	(36)	(39)	(29)	(100)	(54)
−5	−291	−146	−340	247	−26	−236	−234	−357	−73	−168
	(43)	(35)	(30)	(75)	(49)	(50)	(44)	(32)	(127)	(62)
−6	−333	−158	−268	203	73	−203	−221	−319	360	−26
	(43)	(37)	(34)	(78)	(55)	(56)	(51)	(39)	(207)	(82)
−7	−149	−114	−255	93	−78	−101	−256	−318	49	−298
	(87)	(50)	(39)	(90)	(84)	(68)	(59)	(43)	(370)	(103)
−8	−148	−100	−173	103	47	−70	−291	−298	−564	−53
	(76)	(54)	(43)	(204)	(94)	(75)	(63)	(45)	(245)	(135)
−9	−84	−37	−16	−27	89	−134	−311	−258	−24	558
	(79)	(78)	(59)	(413)	(111)	(90)	(72)	(52)	(297)	(456)
−10	−191	−53	−100	−1,753	−178	−71	−76	−174	436	−161
	(81)	(75)	(76)	(945)	(123)	(99)	(118)	(75)	(682)	(297)
No. of quarters	106,185	107,606	110,576	19,637	22,194	78,640	76,915	83,535	15,785	21,099

Note.—Standard errors are in parentheses. Missouri analyses are based on a 25% random sample of all individuals; North Carolina analyses are based on a 20% sample. Dependent variable is earnings in quarters prior to participation. Comparison cases matched exactly on outcome quarter and, for cols. 4, 5, 9, and 10, also on quarter of participation. Within an exact match subsample, comparison matches are based on radius matching using the propensity score. See text for details. Variables used for propensity score prediction are calendar quarter of the outcome; county unemployment rate during the quarter; dummies identifying 17 county groups in Missouri and 24 county groups in North Carolina; age and age squared; nonwhite; education less than high school; number of children; age of youngest child; proportion of quarters working; dummies for no work and working all 8 quarters, based on 8 quarters prior to entering welfare; total earnings in the 4 quarters immediately prior to entering welfare; earnings in the fifth through eighth quarters prior to entering welfare; and the proportion of the 8 quarters prior to entering welfare in which welfare was received.

parison group is more likely to provide adequate estimates of program impact, but the possibility of bias remains.

The results for North Carolina for job search/readiness training, when compared with assessment (table 5, col. 9) show a negative impact of $117 in the quarter preceding service, but otherwise estimates are inconsistent and generally not statistically significant. In contrast, the estimates of intensive training impacts on prior earnings for North Carolina (reported in col. 10) are almost all negative, and most are statistically significant. The average across the 6 quarters prior to service is −$134. This suggests that individuals participating in intensive activities are systematically less successful than those selected into assessment, even when individuals' measured characteristics and labor market success prior to welfare entry are controlled.

We also performed specification tests for the fixed-effects model. Given that the results reported in table 5 imply that cases receiving no services differ dramatically in terms of prior measures from those that receive services, we limit consideration to estimates that use assessment as the comparison group. As in the previous specification test, we identify an "effect" on earnings prior to receipt of actual service. As detailed above, the fixed-effects specification test requires that a quarter be chosen as a "pseudo" service quarter prior to the outcome quarter of interest. Estimates reported here are based on all such possible quarters.[27]

Table 6 suggests that the fixed-effects model is more successful than the simple levels structure. In the case of job search/readiness, estimates are not statistically significant for either state. In the case of intensive activities in Missouri (col. 2), although the estimate for the immediately prior quarter borders on statistical significance, it is of modest size; other estimates are small and not statistically significant. However, the results provide only limited support for the use of the fixed-effects estimate for intensive activities in North Carolina. The estimates of the effect of participation on earnings in the prior 2 quarters are −$165 and −$99, which, although smaller than comparable estimates in table 5, are statistically significant. Impact estimates for earnings in earlier quarters are much smaller and not statistically significant, in marked contrast to results reported in table 5, suggesting that the fixed-effects model is more likely to be successful in providing valid impact estimates.

VII. Effects of Work Component Participation

Given the results of the specification tests, we focus on models that obtain estimates based on a comparison of individuals who receive job

[27] A given outcome quarter will therefore appear multiple times in the analysis. In calculating estimates, we have weighted each quarter by the inverse of the number of times it appears in the analysis. Standard errors also are calculated using this weighting scheme.

Table 6
Matching Estimates of "Effects" of Services on Earnings in Prior Quarters: Fixed-Effects Estimates

	Missouri		North Carolina	
Quarter Relative to Service	Job Search/ Readiness vs. Assessment (1)	Intensive vs. Assessment (2)	Job Search/ Readiness vs. Assessment (3)	Intensive vs. Assessment (4)
−1	57	−41	−46	−165
	(57)	(21)	(38)	(40)
−2	25	−17	−21	−99
	(39)	(19)	(49)	(36)
−3	19	−4	37	−24
	(46)	(20)	(71)	(33)
−4	−4	−13	−1	−16
	(79)	(21)	(78)	(42)
−5	14	−7	54	−21
	(88)	(25)	(121)	(57)
−6	34	8	−6	0
	(50)	(34)	(173)	(66)
−7	51	−13	−64	−1
	(123)	(51)	(211)	(70)
−8	17	−12	−7	−81
	(292)	(56)	(173)	(140)
−9	2	−2	35	−60
	(369)	(62)	(58)	(288)
No. of quarters	9,859	11,728	6,506	10,370

Note.—Standard errors are in parentheses. Missouri analyses are based on a 25% sample of all individuals; North Carolina analyses are based on a 20% sample. Dependent variable is earnings in quarters prior to participation minus earnings prior to "pseudo" service quarter. Comparison cases matched exactly on outcome quarter, quarter of participation, and "pseudo" service quarter. Within an exact match subsample, comparison matches are based on radius matching using the propensity score. See text for details. See table 5 for the variables used for the propensity score prediction.

search/readiness training and intensive services with those receiving assessment services. Estimates based on the simple outcome model are reported in table 7 and figure 3, and estimates based on the fixed-effects model are presented in table 8. In each case, we examine estimates of impact on earnings in the quarter of participation and the 15 successive quarters (14 for the fixed-effects model) in Missouri and North Carolina.

Panel A in figure 3 shows that the estimated effects for job search/readiness do not change dramatically at greater time lags. In both states, effect estimates are slightly lower in the quarter of service and the following quarter but after an increase do not change up through quarter 13. However, the levels of effect estimates differ dramatically by state, with numbers very close to zero in Missouri and consistently negative in North Carolina. Notwithstanding these differences, we have no evidence that this work component provides any meaningful benefits—either in the long term or the short term—for participants. Standard errors, illustrated with bars in figure 3, are small, suggesting that observed patterns are not influenced by sampling error.

Table 7
Matching Estimates of Impact of Services on Earnings in Later Quarters

	Missouri		North Carolina	
Quarter Relative to Service	Job Search/ Readiness vs. Assessment (1)	Intensive vs. Assessment (2)	Job Search/ Readiness vs. Assessment (3)	Intensive vs. Assessment (4)
0	−46	−302	−245	−383
	(9)	(6)	(8)	(6)
1	−52	−444	−210	−505
	(13)	(8)	(13)	(9)
2	12	−303	−124	−279
	(15)	(10)	(18)	(13)
3	8	−238	−265	−263
	(16)	(10)	(21)	(15)
4	33	−137	−143	−222
	(17)	(11)	(18)	(13)
5	−5	−115	−172	−204
	(18)	(12)	(22)	(15)
6	6	−23	−159	−166
	(20)	(13)	(21)	(16)
7	14	17	−146	−151
	(21)	(14)	(22)	(18)
8	−20	57	−190	−119
	(22)	(15)	(24)	(19)
9	−5	67	−182	−107
	(24)	(17)	(27)	(22)
10	73	137	−202	−66
	(27)	(19)	(30)	(30)
11	40	136	−219	−9
	(32)	(22)	(34)	(35)
12	73	174	−218	−30
	(39)	(28)	(41)	(39)
13	13	162	−95	109
	(47)	(36)	(49)	(45)
14	168	355	−113	60
	(59)	(44)	(59)	(57)
15	174	402	−126	105
	(101)	(78)	(104)	(121)
11–15 (mean)	94	246	−154	47
	(27)	(21)	(28)	(30)
No. of quarters	234,251	265,317	257,002	311,482

NOTE.—Standard errors are in parentheses. Dependent variable is earnings in quarters following program participation. Comparison cases matched exactly on outcome quarter and quarter of participation. Within an exact match subsample, comparison matches are based on radius matching using propensity score. See text for details. See table 5 for the variables used for the propensity score prediction.

Panel B in figure 3 presents impact estimates for intensive training. For Missouri, estimates imply reduced earnings in the quarter of service and the immediately following quarter but progressively greater earnings in later quarters, with impact becoming positive in quarter 7. In contrast, although the pattern of effects is very similar, the North Carolina effects are below those of Missouri, especially in the later quarters. Only beginning in quarter 13 are the North Carolina impact estimates positive. The average impact in Missouri for the last 5 quarters is nearly $250 per year

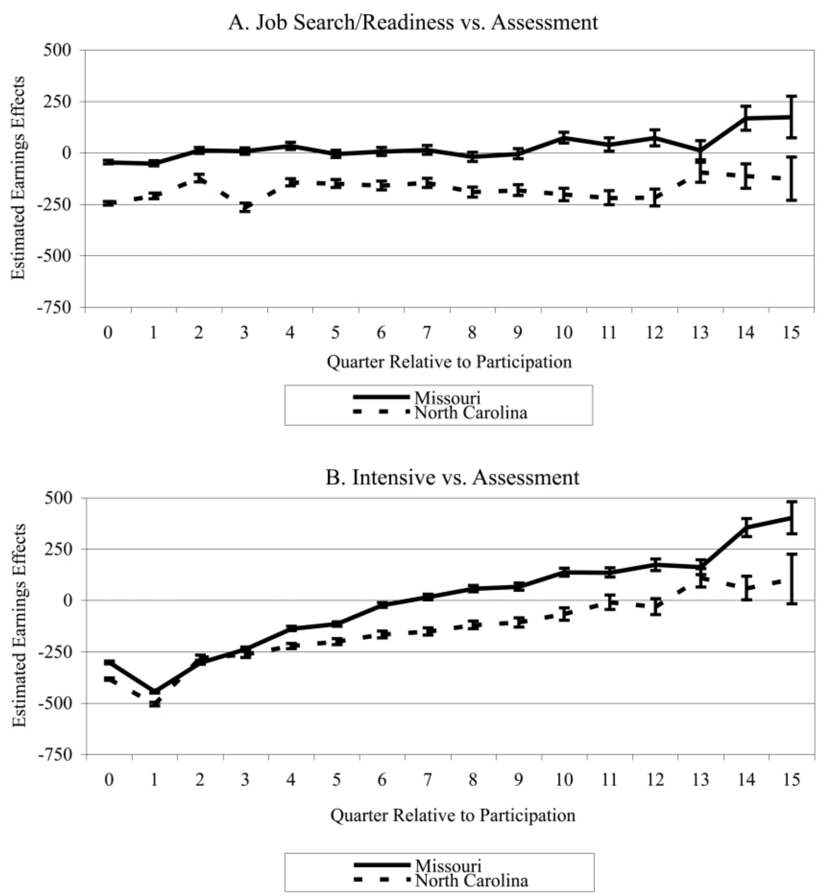

FIG. 3.—Patterns of estimated impact by quarter relative to participation

and clearly statistically significant. For North Carolina, the average is less than $50 and is not statistically significant (see table 7).

Our specification tests suggested that estimates based on the comparison of services with assessment may not fully control for selection into service type. The bias would appear to be particularly serious for North Carolina. In the quarter prior to participating in intensive services, individual earnings are $184 less than the earnings of those who are only assessed, and earnings in the prior 3 quarters are between $66 and $209 below the earnings of those who are merely assessed (see table 5). Bearing this in mind, we may well question the estimates obtained from this specification.

The results of the specification test for Missouri are more supportive of this model, and we therefore believe the Missouri estimates are better indicators of program impact. These estimates imply that the program

Table 8
Estimates of Impact of Services on Earnings in Later Quarters: Fixed-Effects Estimates

Quarter Relative to Service	Missouri: Job Search/ Readiness vs. Assessment (1)	Missouri: Intensive vs. Assessment (2)	North Carolina: Job Search/ Readiness vs. Assessment (3)	North Carolina: Intensive vs. Assessment (4)
0	−115	−337	−248	−283
	(12)	(6)	(11)	(7)
1	−81	−436	−176	−358
	(16)	(9)	(16)	(10)
2	−24	−316	−141	−184
	(18)	(11)	(23)	(14)
3	12	−265	−120	−118
	(20)	(11)	(24)	(16)
4	22	−161	−106	−80
	(22)	(12)	(23)	(15)
5	−31	−153	−107	−64
	(23)	(13)	(24)	(17)
6	−26	−65	−101	−21
	(26)	(14)	(26)	(18)
7	11	−39	−68	13
	(28)	(15)	(29)	(20)
8	−50	−16	−89	39
	(31)	(17)	(31)	(22)
9	18	7	−109	60
	(34)	(20)	(34)	(25)
10	117	97	−128	97
	(38)	(22)	(38)	(31)
11	138	133	−148	152
	(46)	(26)	(44)	(37)
12	163	127	−145	116
	(62)	(34)	(52)	(45)
13	111	144	18	313
	(75)	(44)	(63)	(53)
14	328	363	−68	235
	(104)	(57)	(76)	(68)
11–14 (mean)	185	191	−86	204
	(37)	(21)	(30)	(26)
No. of quarters	158,285	193,178	177,064	236,162

Note.—Standard errors are in parentheses. Dependent variable is earnings for quarters following participation minus average earnings for quarters prior to participation. Participation in quarter 0 is not considered, because prior earnings following welfare entry are not available. Thus, only 14 outcome quarters are available following participation. Comparison cases matched exactly on outcome quarter and quarter of participation. Within an exact match subsample, comparison matches are based on radius matching using propensity score. See text for details. See table 5 for the variables used for the propensity score prediction.

has some costs, but the overall net returns would appear to be positive. In terms of forgone earnings, if we simply add up the negative earnings increments estimated for quarters 0–6 in table 7, we find that total earnings in these 7 quarters are reduced by $1,562 (standard error of $27). The average earnings increment in the last 5 quarters is $246, or nearly $1,000 per year. If such an increment remained in effect for a 30-year working life, using a 6% discount rate, the earnings gains would exceed the initial earnings loss by more than $10,000 in net present value at the time of participation. Since most estimates suggest that the costs of providing training programs are often less than $2,000 per person (e.g., Orr et al. 1996), these figures obviously suggest that the program would pay for itself.

Although the results differ in important respects, it is worth noting that in both states the basic pattern of earnings movement is similar. Individuals receiving job search services experience very little earnings gains after participating in the program relative to those who are only assessed. In contrast, in both states individuals receiving intensive services experience fairly steadily rising earnings after receiving training, relative to those who are only assessed. These patterns closely mirror the results in Hotz et al. (2006, in this issue). In their revaluation of the California GAIN program Hotz et al. find that, relative to persons receiving job search services, individuals receiving more intensive training exhibit rising earnings in the period after completing the program and that, in the long run, the more intensive training programs provide greater benefits than the short-term job search programs. This is clearly what we see in Missouri—in the long run the intensive training programs produce higher benefits than the short-term job search programs. This is also the case in North Carolina, although our estimates suggest that the benefits for job search are negative, while those for the more intensive programs are small.

Given that our specification tests suggest potential problems with the model, especially for North Carolina, table 8 presents estimates based on the differenced analysis that controls for individual fixed effects that are constant over time.[28] For Missouri, we find that estimates of the effect of intensive services change very little relative to the previous model, whereas estimates for job search/readiness shift. Job search/readiness appears to have a moderate negative effect on earnings during the earlier period and a substantial positive effect later. If we believe these estimates are valid, we would judge job search/readiness to provide benefits as great as the intensive services. Estimates in North Carolina shift most dramatically for intensive services. The pattern of returns now mimics that for Missouri fairly closely. Job search/readiness continues to have a substantial negative impact, but not as great as the negative effect estimated by the prior model.

In summary, given the specification test results, we believe that the estimates reported in either table 7 or table 8 for Missouri provide useful estimates of program impact, indicating that in the long run, the more

[28] In order to estimate the fixed-effects model, it is necessary to have at least 1 quarter prior to service receipt after coming onto welfare. Hence, the estimates do not include those receiving services in quarter 0. Since individuals are only followed for a total of 16 quarters, table 8 reports estimates only for the 14 quarters after the participation quarter.

intensive training activities produce greater benefits than short-term job search/readiness. In the case of North Carolina, we put greater emphasis on the estimates in table 8, which show a similar pattern of effects.

VIII. Conclusion

We use administrative data on welfare participants in two states, Missouri and North Carolina; earnings data for a number of periods after these individuals have participated in a work component; and propensity score matching methods to estimate the effect on earnings of participating in short-term job search/readiness activities versus participating in longer-term intensive training. Our results, taken as a whole, indicate that short-term job search/readiness programs have minimal long-term impacts. In contrast, we find that the longer-term intensive training programs initially have substantial negative effects, but these effects turn positive within 2 years of program participation and appear to persist.

Our results are similar to those of Hotz et al. (2006, in this issue) and suggest that more intensive training programs provide greater benefits than short-term programs. These results also suggest that an emphasis on work-first programs is likely misplaced and that administrators should place more emphasis on programs designed to enhance participants' general human capital.

Our analysis also underscores the point that in order to judge the efficacy of long-run intensive training programs, participants must be followed for an extended period after their program involvement, since it takes a considerable amount of time before effects become positive. Our estimates are consistent with the pattern of returns expected from most human capital investments—for example, college education—in that they often require some period of time to appear. Unfortunately, the policy emphasis on work-first activities reflects in part the fact that performance measures used in evaluating welfare and other public training programs are short-term in focus, and it is clear from this study that a short-term perspective will underestimate the benefits of more intensive training and possibly misdirect the allocation of training resources.

Finally, our results, like those of Autor and Houseman (2005), also illustrate the importance of applying tests to evaluate the underlying assumptions for models used to estimate program effects based on nonexperimental data. In the absence of specification tests, researchers are left with little basis for choosing among results produced by alternative models.

Appendix

I. Coding Methodology and Source for Variables in X_{it}

Time-Varying Characteristics in X_{it}

Variable: Calendar-quarter indicators for outcome quarter.
Coding: Indicator equals one if current quarter and equals zero otherwise (18 quarters).
Source: Generated from welfare and unemployment insurance records.

Variable: Age and age squared.
Coding: Age in years at end of current quarter, rounded up to the nearest twelfth of a year.
Source: Calculated using the date of birth recorded in individual's welfare records.

Variable: Age of youngest child.
Coding: Age of youngest child in years at the end of the current quarter. Birth date of the youngest child is established as of the most recent quarter receiving welfare. Rounded up to the nearest twelfth of a year.
Source: Date of birth of youngest dependent child associated with an individual's welfare case.

Variable: Number of children.
Coding: Number of children as of the most recent quarter receiving welfare.
Source: Number of dependent children associated with an individual's welfare case.

Variable: County unemployment rate.
Coding: Mean monthly unemployment rate for current quarter in most recent county of residence while receiving welfare.
Source: State unemployment statistics.

Variable: Workforce Investment Area (WIA) geographic indicators.
Coding: Indicator for a WIA equals one if an individual's county of residence as of the most recent quarter receiving welfare is in the given WIA and zero otherwise. In Missouri, which had fewer WIA areas, separate indicator variables for counties with a population greater than 100,000 were created, even if the county was part of a larger WIA area. There were 17 geographic areas identified in Missouri and 24 in North Carolina.
Source: U.S. Department of Labor WIA classifications.

Time-Invariant Individual Characteristics in X_{it} (May Vary by Spell for an Individual)

Variable: Race indicator.
Coding: Equals one if individual is not Caucasian; equals zero otherwise.
Source: Individual welfare records.

Variable: Education indicator.
Coding: Equals one if an individual had not earned a high school diploma or equivalent as of the first quarter of current welfare spell; equals zero otherwise.
Source: Education records associated with individual's welfare case.

Variable: Proportion of 8 quarters prior to welfare entry receiving welfare.
Coding: Range from zero (no receipt of welfare in 8 quarters prior to welfare entry) to 0.875 (received welfare payments in 7 of 8 prior quarters). Note: by definition, an individual is coded as entering welfare in quarter t only if she received no welfare in quarter $t - 1$.
Source: Individual welfare records.

Variable: Proportion of 8 quarters prior to welfare entry with positive earnings.
Coding: Range from zero (no labor market earnings) to one (positive earnings in all 8 quarters). Variables in X_{it} also include an indicator that equals one if and only if the proportion working variable equals zero and an indicator that equals one if and only if the proportion working variable equals one.
Source: State unemployment insurance wage records.

Variable: Cumulative earnings in the 4 quarters preceding the welfare entry.
Coding: Missing earnings data coded as zero.
Source: State unemployment insurance wage records.

Variable: Cumulative earnings in quarters 5–8 prior to welfare entry.
Coding: Missing earnings data coded as zero.
Source: State unemployment insurance records.

II. Bias Adjustment

Our estimates of impact include a bias adjustment term, following Abadie and Imbens (2006), which corrects for any differences between treated and matched comparison cases. Consider the impact estimate as specified in the article as

$$\hat{\delta}^{s}_{itd} = Y^{s}_{itd} - Y^{0}_{t'}(X_{it'}),$$

where Y^{s}_{itd} is earnings in quarter $t' = t + d$ for treated case i receiving service s in quarter t. The bias adjustment implies that the estimate of earnings based on nonparticipant comparison cases matched to treated case i is given as

$$Y^{0}_{t'}(X_{it'}) = (1/N_i) \sum_{j \in \{j \mid r \geq |P(X_{jt'}) - P(X_{it'})|\}} Y^{0}_{jt'} - (X_{jt'} - X_{it'})\beta,$$

where N_i is the number of comparison cases matched to case i, $Y^{0}_{jt'}$ is the observed earnings of comparison case j matched to case i, $X_{jt'}$ is a vector of observed characteristics of comparison case j, r is the matching radius, $P(X)$ is the propensity score, and β is the coefficient estimated from a regression predicting earnings within the weighted comparison cases in the treatment subsample.

For estimates based on comparison of assessment with more intensive services, the impact estimate for case i is written as

$$\hat{\delta}^{s}_{itd} = Y^{s}_{itd} - Y^{1}_{td}(X_{it'}).$$

We may write the bias-adjusted estimate based on the comparison cases matched to i as

$$Y^{1}_{td}(X_{it'}) = (1/N_i) \sum_{j \in \{j \mid r \geq |P(X_{jt'}) - P(X_{it'})|\}} Y^{1}_{jtd} - (X_{jt'} - X_{it'})\beta.$$

Expressions for bias-adjusted fixed-effects estimates are computed similarly.

Note that if the matching is successful, the expression $(X_{jt'} - X_{it'})$ will be small so that the adjustment will be small. Conversely, if the linear model is fully adequate as a predictor of earnings, the estimates will be consistent even when matching is poor.

III. Analytical Standard Errors

Analytical standard errors reported in the article are based on the assumption that each estimate $\hat{\delta}^{s}_{itd}$ can be treated as an independent draw across all treated cases i. This approach can be interpreted in the following terms. Consider the true impact for individual i based on the model in the article,

$$Y^{s}_{itd} - Y^{0}_{it'} = \delta^{s}_{td}(X_{it'}) + \varepsilon^{s}_{itd}.$$

If we simplify the structure so that the program effect does not depend on $X_{it'}$, we may write

$$Y^s_{itd} - Y^0_{it'} = \delta^s_{td} + \varepsilon^s_{itd}.$$

Of course, we do not observe $Y^0_{it'}$ but rather an estimate of it based on the cases that are matched to i, which we have designated $Y^0_{t'}(X_{it'})$. If we assume that this estimate differs from the true value by an independent estimation error u_{itd} (i.e., $Y^0_{t'}(X_{it'}) = Y^0_{it'} + u_{itd}$), the estimate of the effect for case i can be written as

$$\hat{\delta}^s_{itd} = Y^s_{itd} - Y^0_{t'}(X_{it'}) = \delta^s_{td} + \varepsilon_{itd} + u_{itd}.$$

Based on the assumption that ε_{itd} and u_{itd} are independently and identically distributed, both with mean zero, the standard error of the estimated effect, $\hat{\delta}^s_{td} = (1/N_{td}) \sum_{i \in T_{td}} \hat{\delta}^s_{itd}$, is calculated in the conventional way, as

$$\mathrm{SE}(\hat{\delta}^s_{td}) = \mathrm{SD}(\hat{\delta}^s_{itd})/\sqrt{N_{td}}.$$

The standard error for $\hat{\delta}^s_d$ is calculated using the standard deviation across all $\hat{\delta}^s_{itd}$, where $0 \leq t \leq 10$.

We expect that true standard errors will exceed our estimated standard errors insofar as u_{itd} is not independent for all treated cases i. However, if the variation in ε_{itd} is large relative to the variation in u_{itd}, the bias in the analytical standard error may be modest even where estimation errors are positively correlated.

References

Abadie, Alberto, and Guido W. Imbens. 2006. Large sample properties of matching estimators for average treatment effects. *Econometrica* 74, no. 1:235–67.

Autor, David, and Susan Houseman. 2005. Do temporary help jobs improve labor market outcomes for low-skilled workers? Evidence from random assignment. Working Paper no. 11743, National Bureau of Economic Research, Cambridge, MA (November).

Barnow, Burt, and Daniel B. Gubits. 2002. Review of recent pilot, demonstration, research, and evaluation initiatives to assist in the implementation of programs under the Workforce Investment Act. In *Strategic plan for pilots, demonstrations, research, and evaluations, 2002–2007*, chap. 5. Washington, DC: U.S. Department of Labor, Employment and Training Administration.

Black, Daniel A., and Jeffrey A. Smith. 2004. How robust is the evidence on the effects of college quality? Evidence from matching. *Journal of Econometrics* 121 (July–August): 99–124.

Bloom, Dan, and Charles Michalopoulos. 2001. *How welfare and work* *policies affect employment and income: A synthesis of research*. New York: Manpower Demonstration Research Corp.

Courty, Pascal, and Gerald Marschke. 2004. An empirical investigation of dysfunctional responses to explicit performance incentives. *Journal of Labor Economics* 22, no. 1:23–56.

Dyke, Andrew, Carolyn J. Heinrich, Peter R. Mueser, and Kenneth R. Troske. 2005. The effects of welfare-to-work program activities on labor market outcomes. Discussion Paper no. 1520, Institute for the Study of Labor, Bonn.

Fitzenberg, Bernd, Robert Völter, and Aderonke Osikominu. 2006. Get training or wait? Long-run employment effects of training programs for the unemployed in West Germany. Discussion Paper no. 2121, IZA, Bonn (May).

Friedlander, Daniel, and Gary Burtless. 1995. *Five years after: The long-term effects of welfare-to-work programs*. New York: Russell Sage.

Greenberg, David, Andreas Cebulla, and Stacy Bouchet. 2005. Report on a meta-analysis of welfare-to-work programs. Discussion Paper no. 1312-05, Institute for Research on Poverty, Madison, WI.

Grogger, Jeffrey, and Lynn A. Karoly. 2005. *Welfare reform: Effects of a decade of change*. Cambridge, MA: Harvard University Press.

Grogger, Jeffrey, Lynn A. Karoly, and Jacob Alex Klerman. 2002. Consequences of welfare reform: A research synthesis. Report DRU-2676-DHHS, RAND Corp., Santa Monica, CA.

Gueron, Judith, and Edward Pauly. 1991. *From welfare to work*. New York: Russell Sage.

Haskins, Ron, and Rebecca M. Blank. 2001. Welfare reform reauthorization. *Poverty Research News* 5, no. 6. http://www.jcpr.org/newsletters/vol5_no6/index.html.

Heckman, James J., and V. Joseph Hotz. 1989. Choosing among alternative nonexperimental methods for estimating the impact of social programs: The case of manpower training. *Journal of the American Statistical Association* 84, no. 4:862–74.

Hotz, V. Joseph, Guido W. Imbens, and Jacob A. Klerman. 2006. Evaluating the differential effects of alternative welfare-to-work training components: A reanalysis of the California GAIN program. *Journal of Labor Economics* 24, no. 3:521–66.

Jacobson, Louis S., Robert J. LaLonde, and Daniel G. Sullivan. 1994. The returns from classroom training for displaced workers. Working Paper no. 94-27, Federal Reserve Bank, Chicago.

———. 2004. Estimating the returns to community college schooling for displaced workers. Discussion Paper no. 1017, IZA, Bonn.

Jencks, Christopher, and Kathryn Edin. 1992. The real welfare problem. In *Rethinking social policy*, ed. Christopher Jencks, 204–35. Cambridge, MA: Harvard University Press.

Kornfeld, Robert, and Howard S. Bloom. 1999. Measuring program impacts on earnings and employment: Do UI wage reports from employers agree with surveys of individuals? *Journal of Labor Economics* 17, no. 1:168–97.

Leahey, Erin. 2001. A help or hindrance? The impact of job training on the employment status of disadvantaged women. *Evaluation Review* 25, no. 1:29–54.

Lechner, Michel. 2001. Identification and estimation of causal effects of multiple treatments under the conditional independence assumption. In *Econometric evaluation of active labour market policies*, ed. M. Lechner and F. Pfeiffer, 43–58. Heidelberg: Physica.

Luks, Samantha, and Henry E. Brady. 2003. Defining welfare spells: Coping with problems of survey responses and administrative data. *Evaluation Review* 27, no. 4:395–420.

Mueser, Peter, Kenneth R. Troske, and Alexey Gorislavsky. 2005. Using state administrative data to measure program performance. Working Paper no. 05-20, Department of Economics, University of Missouri, Columbia (December).

Orr, Larry L., Howard S. Bloom, Stephen H. Bell, Fred Doolittle, Winston Lin, and George Cave. 1996. *Does training for the disadvantaged work? Evidence from the National JTPA Study*. Washington, DC: Urban Institute.

Riccio, James, Daniel Friedlander, and Stephen Freedman. 1994. *GAIN: Benefits, costs, and three-year impacts of a welfare-to-work program*. New York: Manpower Demonstration Research Corp.

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

Sianesi, Barbara. 2004. An evaluation of the Swedish system of active labor market programs in the 1990s. *Review of Economics and Statistics* 86, no. 1:133–55.

Smith, Jeffrey, and Petra Todd. 2005. Does matching overcome LaLonde's critique of nonexperimental estimators? *Journal of Econometrics* 125: 305–53.