



Evaluation of Employment Coaching for TANF and Related Populations



Evaluation of Employment Coaching for TANF and Related Populations: Technical Supplement to the Evaluation Design Report

Evaluation of Employment Coaching for TANF and Related Populations: Technical Supplement to the Evaluation Design Report

OPRE Report #2021-221 • October 2021

Quinn Moore • Sheena McConnell • Tim Kautz • April Wu

Submitted to:

Office of Planning, Research, and Evaluation
Administration for Children and Families
U.S. Department of Health and Human Services
330 C Street, SW
Washington, DC 20201
Project Officers: Hilary Bruck and Victoria Kabak

Contract/Task Number: HHSP233201500035I / HHSP23337018T

Mathematica Reference Number: 50327

Submitted by:

Mathematica
1100 1st Street, NE
12th Floor
Washington, DC 20002-4221
Telephone: (202) 484-9220
Facsimile: (202) 863-1763

This report is in the public domain. Permission to reproduce is not necessary. Suggested citation: Evaluation of Employment Coaching for TANF and Related Populations: Evaluation Design Report, OPRE Report #2021-221. Washington, DC: Office of Planning, Research, and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services.

This report and other reports sponsored by the Office of Planning, Research, and Evaluation are available at <http://www.acf.hhs.gov/opre>.

Disclaimer: The views expressed in this publication do not necessarily reflect the views or policies of the Office of Planning, Research, and Evaluation, the Administration for Children and Families, or the U.S. Department of Health and Human Services.



Contents

Analysis priority and outcome selection	1
Categorizing analysis priority	2
Assessing robustness of findings within domains	3
Outcomes in the confirmatory analysis	4
Outcomes in the secondary analysis	7
Outcomes in the exploratory analysis	8
Main approach to estimating impacts	12
Multivariate estimation and covariates	12
Statistical significance	15
Treatment of missing data	15
Treatment of missing baseline data	16
Treatment of missing outcome data	16
Approach to secondary analysis	16
Bayesian analysis	17
Overview of the BASIE approach	17
Guidelines to selecting prior information	18
Recommended source of prior information	18
Selecting priors from Pathways	19
Outcome domains and timing	20
Expected sample sizes	20
Presentation of Bayesian results	20
Pooling across programs	21
Analysis of the COVID-19 pandemic	21
Subgroup analysis based on date of follow-up period	23
Impact analysis by calendar date	24
Approach to Exploratory Analysis	25
Subgroup analysis	25
Estimates of the impacts for those who received program services	27
Mediation analysis	27
Robustness checks	28
References	29

Appendix A. Psychometric analysis of outcomes in the goal-setting and self-regulation skill domain.....	31
---	----

TABLES

Table 1. Classifying confirmatory, secondary, and exploratory analysis	2
Table 2. Confirmatory outcomes	5
Table 3. Secondary outcomes.....	7
Table 4. Exploratory outcomes to be examined for all programs unless noted otherwise	9
Table 5. Availability of baseline versions of confirmatory outcomes	13
Table 6. Baseline characteristics with statistically significant differences between program and control group means, by program	14
Table 7. Conventions for describing statistical significance	15
Table 8. Recommendations for using Pathways to define prior information	19
Table 9. Estimated number of outcomes and studies available to form priors for the main analysis	20
Table 10. Example of presentation of Bayesian results.....	21
Table 11. The timing of the COVID-19 pandemic relative to study enrollment, by program	22
Table 12. The timing of the COVID-19 pandemic relative to study enrollment and first follow-up data collection, by program.....	23
Table 13. Subgroups to include in exploratory analysis	26
Table A.1. Goal-related items in the first follow-up survey	32
Table A.2. Items capturing self-regulation skills in the employment context in the first follow-up survey	32
Table A.3. Criteria used for assessing reliability and validity.....	33
Table A.4. Reliability of measures of self-regulation skills.....	35
Table A.5. Model fit statistics of measures of self-regulation skills.....	35
Table A.6. Correlations between measures of self-regulation skills.....	36

FIGURES

Figure 1. Illustrative example of average monthly earnings from administrative records by calendar quarter and research group status	25
--	----

Introduction

This technical report provides additional details on conducting and reporting on the impact analysis of the Evaluation of Employment Coaching for TANF and Related Populations (hereafter Employment Coaching evaluation). It supplements the design report for the evaluation (Moore et al. 2019). The design report described the evaluation design, identified data sources to be collected, and provided an overview of the implementation and impact study plans. The design report also provided details on how interventions were selected for the study and described the four participating employment coaching programs: (1) Family Development and Self-Sufficiency Program (FaDSS); (2) Goal4 It!; (3) LIFT; and (4) MyGoals for Employment Success (MyGoals).

This report specifies the outcomes that will be included in the confirmatory tests of program effectiveness, documents methodological details for estimating impacts, and describes planned secondary and exploratory analyses. These details provide transparency about the evaluation's analytic approach and lay the groundwork for publicly registering the study with Open Science Framework. The report is divided into four main sections in which we describe: (1) analysis priority and outcome selection, (2) the approach to estimating impacts in the confirmatory analysis, (3) the approach to secondary analysis, and (4) the approach to the exploratory analysis. An appendix presents the results of a psychometric analysis that confirms selection of outcomes in the goal-setting and self-regulation skill domain.

Analysis priority and outcome selection

Employment coaching programs might affect a broad range of outcomes related to self-regulation, employment, economic stability, and other domains. However, the risk of finding a statistically significant result by chance, rather than one representing a true effect of the program, increases with the number of outcomes tested (Schochet 2009). Therefore, we must balance the need to examine the range of outcomes these programs aim to affect with the need to minimize multiple comparison concerns.

We follow two approaches to minimize the risk of focusing on findings that are statistically significant by chance. First, we will restrict the number of outcomes used for determining program effectiveness. The confirmatory tests of program effectiveness will be identified through the categorization of analysis, as described below, and will not be adjusted for multiple comparisons. Before conducting the analyses, we will establish a hierarchy of reporting that places findings into one of three categories: (1) those that must be featured in summaries, such as the executive summary, in addition to the main report; (2) those that must be featured in the main report but are featured in summaries only if they add to understanding of impacts on confirmatory outcomes; and (3) those that must be reported in an appendix, and are reported in the main report or summaries only if they add to our understanding of impacts on confirmatory outcomes. Second, we will conduct multiple comparison adjustments as robustness checks. We plan to use the results that are unadjusted for multiple comparisons for the main analysis, because the statistical adjustments reduce statistical power, or the likelihood of identifying a true effect (Schochet 2009). However, we will assess whether the results for confirmatory analysis are robust to multiple comparison adjustment. We will qualify findings that are not robust to these adjustments as having weaker evidence of effectiveness.

CATEGORIZING ANALYSIS PRIORITY

We plan on using a reporting hierarchy in which analysis is specified in advance as confirmatory, secondary, or exploratory (Table 1). The confirmatory analysis will be used as the core test of program effectiveness. As such, it will be presented prominently in the main report and in summary and bottom-line descriptions of program effectiveness. Secondary analysis includes impacts in domains that are less central to program goals. Secondary findings will be presented in the main report; they will also be included in summary sections of the report if they contribute to understanding the confirmatory findings. All other analysis will be exploratory. Because exploratory findings are not part of the main test of program effectiveness, there is flexibility in what outcomes are included in the analysis and in how they are specified. We will present exploratory findings in appendix tables. In addition, we will include selected exploratory findings in the main report and summary sections of the report if they help us interpret confirmatory findings.

Table 1.
Classifying
confirmatory,
secondary, and
exploratory
analysis

Type of analysis	Required presentation of findings	Permitted presentation of findings
Confirmatory	Overview, executive summary, and main report	All sections of the report
Secondary	Main report	Overview and executive summary only if helpful in interpreting confirmatory analysis
Exploratory	Appendices	Main report and overview and executive summary sections only if helpful in interpreting confirmatory analysis

The text box below illustrates how conclusions would be presented in the report.

This approach has several advantages:

- It ensures that readers will be informed if the program did not generate statistically significant impacts on confirmatory and secondary outcomes, because those findings will be reported regardless of statistical significance.
- It avoids potential criticism that the researchers focused on findings that happened to emerge as statistically significant.
- It allows for flexibility in reporting interesting findings that emerge in the exploratory analysis.
- By presenting all exploratory findings in the appendices, we give readers an honest view of the number of comparisons involved in the exploratory analysis. This will be further underscored by describing the construction of all exploratory measures in the technical appendices.
- It provides an organizing principle for other analyses of interest. For example, we can specify that the subgroup analysis will include only confirmatory outcomes.

Sample conclusions discussed in report sections

- Statistically significant impacts on both a confirmatory earnings outcome and an exploratory educational attainment outcome
 - The main report and summaries would discuss the positive impact on earnings and suggest that this impact could be linked to the positive impact on educational attainment, among other factors. Findings on educational attainment would be identified clearly as part of the exploratory analysis.
 - Example text: The positive impact on earnings could be related to positive impacts on educational attainment. For example, in exploratory analysis, we found that program group members were more likely than control group members to earn a GED during the follow-up period.
- No statistically significant impact on a confirmatory earnings outcome but a statistically significant impact on an exploratory educational attainment outcome
 - The main report and summaries would discuss the fact that the program did not affect earnings. Findings on educational attainment would be included in the main report but with very clear designation of the analysis as exploratory. These exploratory findings on educational attainment would not appear in bottom-line summaries.
 - Example text: Although the program did not affect earnings, we did find positive impacts on some hard skills. For example, in exploratory analysis we found program group members were more likely than control group members to earn a GED during the follow-up period. This pattern could suggest that the intermediate effects on skill were not large enough to translate into effects on earnings, or that the study's follow-up period was not long enough to capture impacts on earnings related to training.
- No statistically significant impact on a confirmatory earnings outcome or an exploratory educational attainment outcome
 - The main report and summaries would discuss the fact that the program did not affect earnings. Findings on educational attainment would be reported in the appendix but not referenced in the main report or summaries.

Assessing robustness of findings within domains

Depending on the number of confirmatory outcomes, the risk of a spurious finding may still be higher than desired. For example, consider a study in which the confirmatory analysis includes 20 outcomes. If the program had no impact on any of the confirmatory outcomes, and all 20 impacts were independent, this would generate a 64 percent chance of finding at least one statistically significant impact by chance. To address this concern, we focus the main test of effectiveness for each domain on a single outcome where possible. As a robustness check on the strength of patterns of statistical significance for domains with multiple outcomes, the team will use conventional statistical adjustments for multiple comparisons for confirmatory outcomes within each domain. These adjustments target an overall significance level within a domain by setting more stringent thresholds (p -values) at which individual statistical tests are considered significant. The team will use the Benjamini-Hochberg method, which takes into account both the number of comparisons and the strength of impacts to determine the thresholds at which p -values are considered statistically significant (Benjamini and Hochberg 1995). We will present unadjusted findings as our main

estimates but will note in the text of the main report and summaries when statistically significant impacts are not robust to adjustments to multiple comparisons within domains. This will enable readers to make informed interpretations of the findings.

Outcomes in the confirmatory analysis

The confirmatory analysis will focus on a small set of outcomes across domains that are central to the programs' goals and are feasible to assess, given the study's sample size and length of follow-up period (Table 2). The confirmatory analysis will include separate estimates of impacts for each program; the secondary analysis will also include impact estimates that pool across the FaDSS, Goal4 It!, LIFT, and MyGoals programs for outcomes included in the confirmatory analysis. In selecting outcomes for the confirmatory analysis, we considered each program's logic model and the outcomes most central to the program's goals. For all programs, the outcome domains for the confirmatory analysis include (1) self-regulation and goal-related skills, (2) labor market outcomes, and (3) economic well-being. For FaDSS and Goal4 It!, the two programs for which TANF participation is an enrollment criterion, the confirmatory analysis also includes a fourth domain: receipt of public assistance.

The evaluation will collect survey and administrative data for study participants at two follow-up points. We selected confirmatory outcomes for the first and second follow-up periods in accordance with the hypothesized timing of program impacts. The data collection for the first follow-up occurs at 9 months after study enrollment for the FaDSS, Goal4 It!, and LIFT programs and at 12 months after study enrollment for the MyGoals programs. Study participants were first asked to complete the survey at 9 months (FaDSS, LIFT, and Goal4 It!) or 12 months (MyGoals sites) after study enrollment. Attempts were made for them to complete the survey over an additional 9 months. On average, study participants completed the survey at the following months after study enrollment: 11.6 months (FaDSS), 11.1 months (Goal4 It!), 10.5 months at (LIFT), and 13.8 months (MyGoals). Study participants in all programs will be asked to complete the second follow-up survey at 21 months after study enrollment.

Table 2.
Confirmatory
outcomes

Outcome, data source, and program	Measure	Justification for selection and other comments
Self-regulation and goal-related skills		
Goal setting and attainment • First and second follow-up survey data <i>All programs</i>	<p>Eight-item scale ($\alpha = 0.86$) indicating the average level of agreement—from “strongly disagree” (= 0) to “strongly agree” (= 3)—that a respondent reports on items about goal-related skills:</p> <ul style="list-style-type: none"> • I know I need to get a job or a better job and really think I should work on finding one • I set employment goals based on what is important to me or my family • I set long-term employment goals that I hope to achieve (such as finding a job, finding a better job, getting promoted, or enrolling in further education) • I set specific short-term goals that will allow me to achieve my long-term employment goals • Based on everything I know about myself, I believe I can achieve my employment goals • When I set employment goals, I think about barriers that might get in my way and make specific plans for overcoming those barriers • Even when I face challenges, I continue to pursue my employment goals • I keep track of my overall progress toward my long-term employment goals and adjust my plans if needed <p>These are study-developed questions. This measure is available on both the first and second follow-up surveys.</p>	<p>“Goal setting and attainment” is the centerpiece both of employment coaching generally and the programs participating in the evaluation specifically. All programs in the evaluation intend to improve goal setting and attainment as participants receive program services and for these improvements to persist over time. We therefore include this measure in the confirmatory analysis for both the first and second follow-up periods. The psychometric properties of this measure indicate that it is appropriate for the evaluation populations, as described in the appendix.</p> <p>Other aspects of self-regulation skills measured in the survey, such as task management, might be improved through goal-setting. These skills are not explicitly targeted by FaDSS or LIFT. One or more such skills might be a focus for some MyGoals and Goal4 It! participants, as needed. However, these programs do not set out to improve any specific skill. Thus, we propose examining other self-regulation skills as part of the exploratory analysis rather than the main test of program effectiveness in this domain.</p>
Labor market outcomes		
Earnings • First and second follow-up survey data <i>All programs</i>	<p>Average monthly earnings during the follow-up period. For the first follow-up period, we will define the reference period for the measure as the following:</p> <ul style="list-style-type: none"> • The first nine months after study enrollment for FaDSS, Goal4 It!, and LIFT • The first twelve months after study enrollment for the MyGoals programs <p>For the second follow-up period, we will define the reference period as follows:</p> <ul style="list-style-type: none"> • Months 10 through 21 after study enrollment for FaDSS, Goal4 It!, and LIFT • Months 13 through 21 after study enrollment for the MyGoals programs 	<p>All programs in the evaluation aim to improve labor market outcomes and would expect these improvements to emerge by the time of the first follow-up and persist over time. For this reason, we include both first and second follow-up measures in the confirmatory analysis.</p> <p>We selected earnings for the confirmatory analysis of labor market outcomes because it encompasses a wide range of ways in which the interventions could affect the labor market success of participants. These include increasing their likelihood of working at all, working more regularly (more weeks, months, or quarters), working more hours when they do work (full time instead of part time, for example), and earning higher wages when they do work. All these effects would show up as an increase in total earnings.</p>

Table 2.
Confirmatory
outcomes
(continued)

Outcome, data source, and program	Measure	Justification for selection and other comments
Labor market outcomes		
<p>Earnings</p> <ul style="list-style-type: none"> • First and second follow-up period administrative data <p><i>FaDSS, Goal4 It!, and MyGoals</i></p>	<p>Average monthly earnings during the follow-up period. For the first follow-up period, we will define the reference period for the measure as follows:</p> <ul style="list-style-type: none"> • The first three quarters after study enrollment for FaDSS and Goal4 It! • The first four quarters after study enrollment for the MyGoals programs <p>For the second follow-up period, we will define the reference period as follows:</p> <ul style="list-style-type: none"> • Quarters 4 through 7 after study enrollment for FaDSS and Goal4 It! • Quarters 5 through 7 after study enrollment for the MyGoals programs 	<p>We recommend examining earnings using both administrative records data (NDNH data) and survey data because these data sources have both strengths and limitations. Unlike survey data, NDNH data have no recall error, and they are available for a longer reference period than the follow-up survey. In contrast, the main advantage of survey data is that they will pick up additional kinds of employment that are not covered by administrative data. Survey data cover self-employment (such as Uber driving and gig economy employment) and under-the-table or informal employment, which can be common among low-wage workers and are often not covered by unemployment insurance benefits and thus not included in NDNH data.</p> <p>By examining earnings from both data sources, we reduce the risk of missing the effect that programs may have had on earnings in either formal or informal jobs. The proposed approach—focusing on earnings and using data from administrative and survey data—has been widely used in prior studies of the impacts of employment and training programs on low-income individuals. Earlier studies that have used this approach include PACT, Parents' Fair Share, Job Corps, WIA Gold Standard Evaluation, and CSPED (see Barnow and Greenberg 2015 for a comprehensive review of many of these earlier studies).</p> <p>We do not include earnings based on administrative records for LIFT participants in the confirmatory analysis because only about half of the LIFT sample provided a valid social security number at the time of study enrollment. As a result, this analysis is at risk of attrition bias and will be included in the exploratory analysis.</p>
Receipt of public assistance		
<p>TANF benefit receipt</p> <ul style="list-style-type: none"> • Second follow-up period administrative data <p><i>FaDSS and Goal4 It! programs</i></p>	<p>Average monthly TANF benefit amount during the second follow-up period. We will define the reference period as months 10 through 21 after study enrollment.</p>	<p>The FaDSS and Goal4 It! programs include TANF benefit receipt among their enrollment criteria. Reducing participation in TANF is a goal of these programs. Because impacts on TANF benefit receipt should emerge after impacts on earnings, TANF benefit receipt for the second follow-up period is included in the confirmatory analysis, but the first follow-up period outcome is not. The first follow-up outcome is included in the exploratory analysis.</p>

Table 2.
Confirmatory
outcomes
(continued)

Outcome, data source, and program	Measure	Justification for selection and other comments
Economic well-being		
Economic hardship • First and second follow-up survey data <i>All programs</i>	A count ranging from 0 to 6 of the number of the following coping strategies used to stretch budgets: • Cut the size of meals or skip meals because couldn't afford enough food • Moved in with other people because of financial problems • Asked to borrow money from friends or family • Went without a phone because could not afford to pay the bill • Sold belongings or took out a payday loan • Went without medical care because of cost	All programs in the evaluation intend to improve the material well-being of their participants. This economic hardship scale reflects the extent to which scarce economic resources affected key aspects of material well-being, such as food, housing, and medical care.

Note: The first follow-up period is the first 9 months after study enrollment for FaDSS, Goal4 It!, and LIFT; it is the first 12 months after study enrollment for MyGoals. The second follow-up period is 21 months after study enrollment for all programs.

CSPED = National Child Support Noncustodial Parent Employment Demonstration; FaDSS = Family Development and Self-Sufficiency; NDNH = National Directory of New Hires; PACT = Parents and Children Together; WIA = Workforce Investment Act.

OUTCOMES IN THE SECONDARY ANALYSIS

In secondary analysis, we will examine impacts on outcomes in domains that are likely to be of interest to readers but were less central to program goals and thus not part of the central test of program effectiveness. This category does not include domains in the confirmatory analysis. For the secondary analysis, we propose outcomes in four domains (Table 3). We propose including these outcomes in the secondary analysis for all programs.

Table 3.
Secondary
outcomes

Outcome, data source, and program	Measure	Justification for selection and other comments
Hard skill acquisition		
Completion of an education program • First and second follow-up survey data <i>All programs</i>	This is a binary variable that equals 1 if respondents reported completing an education program and 0 otherwise. This outcome is available at both first and second follow-ups.	Program impacts on education could lead to impacts on earnings and other outcomes.
Completion of a training program • First and second follow-up survey data <i>All programs</i>	This is a binary variable that equals 1 if respondents reported completing a training program and 0 otherwise. This outcome is available at both first and second follow-ups.	Program impacts on training could lead to impacts on earnings and other outcomes.

Table 3.
Secondary
outcomes
(continued)

Outcome, data source, and program	Measure	Justification for selection and other comments
Job quality		
Employment in jobs offering fringe benefits • First and second follow-up survey data <i>All programs</i>	This is a binary variable that equals 1 if respondents reported being employed in a job offering fringe benefits at the end of the follow-up periods and 0 otherwise. The reference period for these outcomes will align with the earnings measures in Table 2.	Understanding impacts on job quality is key to understanding impacts on earnings. Moreover, improved job quality could improve other aspects of economic stability and well-being. Receipt of fringe benefits is an important measure of the quality of a job.
Employment challenges		
Specific challenges that impeded employment • First and second follow-up survey data <i>All programs</i>	Seven separate outcomes correspond to a particular challenge: The first is whether (yes or no) the survey respondent reported having a current driver's license at the time of the follow-up survey. The other six outcomes are binary variables indicating whether respondents reported that having the challenge made it very hard or extremely hard for them to find and keep a good job at the time of the follow-up survey: • Child care • Transportation • Right clothes or tools • Right skills or education • Criminal record • Limiting health reason	Program services and potential impacts on self-regulation skills could influence participants' ability to address employment challenges. Such effects could be important for understanding impacts on earnings and receipt of public assistance.
Housing stability		
Unstable housing • First and second follow-up survey data <i>All programs</i>	This binary variable indicates whether respondents were homeless, living in a shelter, or living rent-free at the time of the follow-up surveys.	Increasing housing stability is often a common goal for participants in some of the programs participating in the evaluation. Housing stability could also be influenced by potential impacts on labor market outcomes.

OUTCOMES IN THE EXPLORATORY ANALYSIS

In exploratory analysis, we will examine a broader set of outcomes, which will not be considered key indicators of program effectiveness but could be used to broaden our understanding of overall program effects. The main function of this analysis is to further investigate the confirmatory findings and identify how they emerged.

We provide a list of outcomes to be included in the exploratory analysis below (Table 4). These include supplementary outcomes in the domains with confirmatory outcomes, as well as outcomes from the following domains: (1) education and training receipt, (2) job search, (3) labor market participation, (3) criminal activity, and (4) marital status. In addition, the exploratory analysis will include impact estimates on service receipt outcomes drawn from survey reports collected from both program and control group members. These findings will inform interpretation of other findings because impacts

on these outcomes would supply information about the contrast between the services received by the program and control groups. We will examine these outcomes for all programs, except for outcomes that are in the confirmatory analysis for some programs or not available for some programs, as noted in the table. The list of exploratory outcomes is likely to expand as we further investigate patterns in the confirmatory analysis.

Table 4.
Exploratory
outcomes to be
examined for all
programs unless
noted otherwise

Outcome	Data source
Service receipt	
Number of times received and total time spent in one-on-one employment services (average and frequency) during the first and second follow-up periods	First and second follow-up surveys
Number of times received group employment services (average and frequency) during the first and second follow-up periods	First and second follow-up surveys
Received job assistance focused on setting long-term goals during the first and second follow-up periods	First and second follow-up surveys
Received job assistance focused on setting short-term goals during the first and second follow-up periods	First and second follow-up surveys
Received job assistance focused on planning to achieve your goal during the first and second follow-up periods	First and second follow-up surveys
Received a career assessment during the first and second follow-up periods	First and second follow-up surveys
Received job leads from a program during the first and second follow-up periods	First and second follow-up surveys
Received child care services during the first and second follow-up periods	First and second follow-up surveys
Received transportation assistance during the first and second follow-up periods	First and second follow-up surveys
Received clothes, uniforms, tools or other supplies and equipment during the first and second follow-up periods	First and second follow-up surveys
Received tuition assistance during the first and second follow-up periods	First and second follow-up surveys
Received assistance finding stable housing during the first and second follow-up periods	First and second follow-up surveys
Received assistance with budgeting, credit, banking, or other financial matters during the first and second follow-up periods	First and second follow-up surveys
Received assistance expunging a criminal record or other legal assistance during the first and second follow-up periods	First and second follow-up surveys
Received help related to domestic violence during the first and second follow-up periods	First and second follow-up surveys
Received help with marital and other family relationships during the first and second follow-up periods	First and second follow-up surveys
Received help with child behavioral issues during the first and second follow-up periods	First and second follow-up surveys
Received cash or a gift card during the first and second follow-up periods	First and second follow-up surveys
Received substance use counseling during the first and second follow-up periods	First and second follow-up surveys
Received mental health treatment during the first and second follow-up periods	First and second follow-up surveys

Table 4.
Exploratory
outcomes to be
examined for all
programs unless
noted otherwise
(continued)

Outcome	Data source
Goal setting and self-regulation skills	
Task monitoring, planning, and initiation at the time of the first and second follow-up surveys	First and second follow-up surveys
Emotional control and self-monitoring at the time of the first and second follow-up surveys	First and second follow-up surveys
Employment self-regulation at the time of the first and second follow-up surveys	First and second follow-up surveys
Self-esteem at the time of the first and second follow-up surveys	First and second follow-up surveys
Set an employment goal at the time of the first and second follow-up surveys	First and second follow-up surveys
Education and training receipt	
Participation in an education program during the first and second follow-up periods	First and second follow-up surveys
Participation in a training program during the first and second follow-up periods	First and second follow-up surveys
Receipt of a diploma or degree from an education program during the first and second follow-up periods	First and second follow-up surveys
Receipt of a certificate, license, or diploma from a training program during the first and second follow-up periods	First and second follow-up surveys
Highest level of education at time of the follow-up surveys	First and second follow-up surveys
Job search	
Number of job search activities conducted since random assignment (updated resume, explored requirements for a job, found child care, looked into training, looked into transportation) during the first and second follow-up periods	First and second follow-up surveys
Number of job offers received when working and number of job offers received when not working during the first and second follow-up periods	First and second follow-up surveys
Intensity of job search (frequency of activities) when working and intensity of job offers received when not working during the first and second follow-up periods	First and second follow-up surveys
Employment and earnings	
Earnings by quarter after study enrollment <i>LIFT only, confirmatory outcome for other programs</i>	NDNH data for the first and second follow-up periods
Whether employed by month after study enrollment	First and second follow-up surveys
Number of months employed	First and second follow-up surveys
Earnings by month after study enrollment	First and second follow-up surveys
Whether employed by quarter after study enrollment	NDNH data for the first and second follow-up periods
Number of quarters employed	NDNH data for the first and second follow-up periods
Earnings by quarter after study enrollment	NDNH data for the first and second follow-up periods

Table 4.
Exploratory
outcomes to be
examined for all
programs unless
noted otherwise
(continued)

Outcome	Data source
Employment and earnings	
Whether a new hire during the first and second follow-up periods	NDNH data for the first and second follow-up periods
Hours worked per week during the first and second follow-up periods	First and second follow-up surveys
Number of months employed in job with wage rate over 25 percentile in the US (about \$14) during the first and second follow-up periods	First and second follow-up surveys
Number of jobs with each of the following benefits: health insurance, paid leave, retirement benefits during the first and second follow-up periods	First and second follow-up surveys
Number of months employed in a (full- or part-time) wage and salary job (excluding seasonal, contract, on-call, and odd-jobs) during the first and second follow-up periods	First and second follow-up surveys
Number of months employed in a full-time job during the first and second follow-up periods	First and second follow-up surveys
Number of months self-employed during the first and second follow-up periods	First and second follow-up surveys
Whether employed in a job with high perceived likelihood of promotion in next 12 months at the time of the follow-up surveys	First and second follow-up surveys
Whether satisfied with job held at the time of the follow-up surveys	First and second follow-up surveys
Labor market participation	
Whether in labor market at time of the follow-up surveys (employed or looking for a job)	First and second follow-up surveys
Whether actively engaged at the time of the follow-up surveys (employed, looking for a job, in school or training, or caring for a family member)	First and second follow-up surveys
Receipt of public assistance or social insurance benefits	
Receipt of any income from public assistance/social insurance programs (TANF, SNAP, UI, SSI, SSDI, WIC, or housing assistance) during the first and second follow-up periods	First and second follow-up surveys
Amount of TANF benefits received by month after study enrollment	Administrative records data for the first and second follow-up periods
Total amount of TANF benefits <i>LIFT and MyGoals; this is a confirmatory outcome for FaDSS and Goal4 It!</i>	Administrative records data for the first and second follow-up periods
Total amount of SNAP benefits received by month after study enrollment <i>FaDSS, LIFT, and MyGoals</i>	Administrative records data for the first and second follow-up periods
Total amount of unemployment insurance benefits received during the first and second follow-up periods	NDNH data for the first and second follow-up periods
Whether received housing assistance <i>MyGoals only</i>	Public housing authority administrative data

Table 4.
Exploratory
outcomes to be
examined for all
programs unless
noted otherwise
(continued)

Outcome	Data source
Criminal activity	
Whether convicted of a crime since study enrollment during the first and second follow-up periods	First and second follow-up surveys
Whether convicted of a felony since study enrollment during the first and second follow-up periods	First and second follow-up surveys
Marital status	
Whether married at time of the follow-up surveys	First and second follow-up surveys

NDNH = National Directory of New Hires; SNAP = Supplemental Nutrition Assistance Program; SSDI = Social Security Disability Insurance; SSI = Supplemental Security Income; TANF = Temporary Assistance for Needy Families; UI = Unemployment Insurance; WIC = Special Supplemental Nutrition Program for Women, Infants, and Children.

Main approach to estimating impacts

The main impact estimates for all outcomes will be based on the evaluation's experimental design. Participants eligible for the coaching services have been randomly assigned to one of two groups: (1) a program group offered coaching services or (2) a control group not offered coaching services. With this design, the research groups should be very similar in terms of their characteristics before receiving the intervention. Our basic analytic approach is to compare the outcomes of members of the program and control groups. Because of random assignment, differences in observed outcomes between the program and control groups large enough to be unlikely to be due to chance can be attributed to the offered employment coaching.

The confirmatory analysis will include separate estimates of impacts for each program. For FaDSS, LIFT, and MyGoals—programs that have more than one geographic location participating in the evaluation—the confirmatory analysis will pool estimates across the locations, as the same approach to employment coaching is being implemented. For MyGoals, the secondary analysis will include separate estimates for each program location (Baltimore and Houston), as the sample size for each location is large enough to enable analyses that will examine any differences in outcomes by location. As discussed further below, the secondary analysis will also consider impact estimates that pool across all four programs for outcomes included in the confirmatory analysis, which will allow us to examine the average effect of coaching programs in the Employment Coaching evaluation. Impact estimates from this analysis will be more precise than the program-specific estimates because of the larger sample size.

MULTIVARIATE ESTIMATION AND COVARIATES

We will estimate the impact of the employment coaching program on each outcome using a multivariate weighted least-squares regression model. There are two reasons for estimating a regression model rather than just using the difference in the average value of the outcome between the program and control groups. First, it will enable us to adjust for any differences in baseline characteristics between the program and control groups that emerge by chance, despite the random assignment design. Second, including covariates in the model that are correlated with the outcome measure will improve the statistical precision of the impact estimates (Orr 1999). These two reasons motivate how we will select covariates to include in the model, as described below.

The baseline data available differ by study program because of differences in the study intake process. The baseline information available for MyGoals study participants is less comprehensive than that for other study participants. Three programs—FaDSS, LIFT, and Goal4 It!—administered a baseline survey developed by the Employment Coaching study team. The two MyGoals programs administered a different baseline instrument at the time of study intake and collected some other baseline data from administrative records from the public housing authorities implementing MyGoals.¹ The MyGoals baseline information includes demographics and earnings prior to study enrollment, but it does not include information on baseline employment challenges or goal-setting and self-regulation skills, as the baseline survey administered to the other three programs does.

We will have access to baseline administrative earnings records through the National Directory of New Hires (NDNH) data. However, this data cannot be exported from the NDNH data system. Therefore, we will only use baseline administrative earnings records as a covariate in impact analysis of outcomes based on NDNH earnings records.

We will select a set of covariates based on likely correlation with program outcomes because highly correlated covariates will increase the precision of the impact estimates. We will include baseline versions of all confirmatory outcomes that are available because baseline and follow-up versions of the same measure are likely to be highly correlated (Table 5). We will use the same set of baseline variables for all outcomes. This will enable us to control for a set of relevant characteristics and will simplify programming the impact estimation.

Table 5.
Availability of
baseline versions
of confirmatory
outcomes

Outcome	Availability for FaDSS, Goal4 It!, and LIFT	Availability for MyGoals programs
Self-regulation and goal-related skills		
Goal-setting and attainment, first and second follow-up surveys	Three of the eight items ^a	No
Labor market outcomes		
Earnings, survey data	Earnings in 30 days before enrollment	Administrative records on earnings at enrollment
Earnings, administrative data	Available only for analysis of outcomes based on NDNH data	Available only for analysis of outcomes based on NDNH data
Receipt of public assistance		
Average monthly TANF benefit, administrative data	Available for FaDSS and Goal4 It!	Yes
Economic well-being		
Economic hardship	No	No

Note: The first follow-up period is the first 9 months after study enrollment for FaDSS, Goal4 It!, and LIFT; it is the first 12 months after study enrollment for MyGoals. The medium-term follow-up period is 21 months after study enrollment for all programs.

^a See Appendix Table 1 for specific items.

¹ Study intake for the MyGoals programs began before the baseline data collection instruments used for the other programs were developed and approved.

In addition, for each program, we will include a set of covariates to control for other baseline characteristics with statistically significant differences between the program and control groups. For FaDSS, Goal4 It!, and LIFT, we tested for average differences between program and control group members in a wide range of baseline characteristics, including age, sex, race and ethnicity, marital status, number of adults with whom the respondent lives, number of children with whom the respondent lives, whether the study participant has a high school or General Education Development (GED) diploma, and challenges to employment the study participant faces. For MyGoals, this list included age, sex, race and ethnicity, and whether the study participant has a high school or GED diploma. As expected, given the random assignment research design, there are few significant differences in baseline characteristics between the program and control groups (Table 6). We will include those characteristics for which there are significant differences as covariates in regressions for all outcomes for that program. Because the characteristics with statistically significant differences will not be the same for all programs, the set of characteristics included in the model will likewise not be the same for all programs.

The covariates will also include a set of indicator variables related to the timing of study enrollment. These covariates will control for common timing effects experienced by enrollment cohorts, such as those related to labor market conditions or other factors. We will construct the indicators to group all sample members whose follow-up period ends March 2020 or later. These follow-up periods include the COVID-19 pandemic; other analysis related to the pandemic are discussed in the secondary analysis section below. The indicators will differ for the MyGoals programs because the enrollment period was substantially longer than for other programs, and the first follow-up period differs. For MyGoals, these indicators will include whether enrollment was (1) February 2017 through February 2018, (2) between March 2018 and February 2019, or (3) March 2019 or later. For FaDSS, Goal4 It!, and LIFT, these indicators will include whether enrollment was (1) June 2018 through May 2019 or (2) June 2019 or later.

Table 6.
Baseline
characteristics
with statistically
significant
differences
between program
and control
group means,
by program

Program	Baseline characteristics to be included as covariates
FaDSS	None
Goal4 It!	None
LIFT	<ul style="list-style-type: none"> • Whether married at study enrollment • Reported criminal history as a challenge to finding and keeping a good job
MyGoals	Baltimore: None Houston: Age

Note: All baseline characteristics with statistically significant differences between the program and control groups for a given program will be included as covariates in the regression models used to estimate that program's impacts. For FaDSS, Goal4 It!, and LIFT, we tested for average differences between program and control group members in a wide range of baseline characteristics: age, sex, race and ethnicity, marital status, number of adults with whom the respondent lives, number of children with whom the respondent lives, whether the participant has a high school or GED diploma, and a set of challenges to employment. For MyGoals, this list included age, sex, race and ethnicity, and whether the participant has a high school diploma or GED.

STATISTICAL SIGNIFICANCE

For all outcomes (confirmatory, secondary, and exploratory), we plan to use standard, frequentist methods to estimate the effects of coaching, and we will report statistical significance based on p -values. This approach will allow us to determine whether the employment coaching program has an effect on an outcome and quantify the magnitude of any effects. We will regard these standard estimates as our main ones, reflecting the bottom-line estimate of program effects on a given outcome. In addition, for confirmatory outcomes, we propose to conduct Bayesian analyses that will provide a complementary interpretation about the *probability* that the employment coaching program has particular effects. This approach is described in a later section.

For our main impact estimates, we will deem impact estimates to be statistically significant if the associated p -value of the estimate falls below 5 percent based on a two-tailed hypothesis test (Table 7). We will also note if the associated p -value falls between 5 and 10 percent, classifying these impacts as statistically significant at the 0.10 level.

To help interpret the magnitude of the impact estimates, we will calculate an effect size for each outcome. We will report effect sizes in the main report for outcomes measured as scales because the magnitude of impacts on these outcomes is not easily interpretable. We will report effect sizes for all outcomes in the appendices. For continuous outcomes, we will calculate the effect size as Hedges' g , which equals the impact estimate from the regression model divided by the unadjusted pooled standard deviation of the outcome for respondents across both program and control groups (Hedges 1981). For binary outcomes, we will calculate the effect size as the Cox index, which equals the log odds ratio divided by the constant 1.65 (Cox 1970).

Table 7.
Conventions
for describing
statistical
significance

p -value of impact estimate	Symbol used to denote p -value	Description of impact estimate
$p < 0.01$	***	Statistically significant
$0.01 \leq p < 0.05$	**	Statistically significant
$0.05 \leq p < 0.10$	*	Statistically significant at the 0.10 level
$p \geq 0.10$	None	Not statistically significant

TREATMENT OF MISSING DATA

We will conduct the analysis to account for the possibility that missing data could introduce bias in the impact estimates and reduce statistical power to detect program impacts. Although all study participants must complete the baseline data collection as part of the study enrollment process, some baseline data could be missing if study participants do not respond to certain items. Follow-up survey data could be missing because study participants do not respond to follow-up surveys or because survey respondents do not answer some survey questions. Data from the NDNH could be missing when study participants are not matched to the administrative records because of missing or inaccurate social security numbers. In addition to these strategies, to assess selection into missing data status, we will compare the baseline characteristics of those who are missing a given type of data and those who are not.

Treatment of missing baseline data

When one or more covariates have missing data, we will use dummy variable adjustment, which involves setting any missing baseline values to a single constant value and including flag variables for missing values as additional covariates in the regression model. This approach is appropriate when the covariates are not correlated with research groups, as is the case in evaluations with a random assignment design (Deke and Puma 2013; Puma et al. 2009).

Treatment of missing outcome data

We will estimate all regressions using weights to account for sample members who did not complete the follow-up survey or could not be matched to the administrative data because of missing or inaccurate social security numbers. The nonresponse weights will adjust the data to be representative of all sample members, not just those who completed the survey or could be matched to an administrative record. Using regression analysis, we will calculate the weights by estimating, for each program separately, the probability of nonresponse for study participants as a function of their baseline characteristics. We will adjust the standard errors of the impact estimates to account for the variability associated with these weights.

We will use imputation to address item nonresponse that affects a subset of items used to create survey outcomes. For example:

- If a sample member responded to at least two-thirds of the items on a scale, we will use the average scale score for that person based on the available items.
- We will impute minor missing items from the job grid with midpoint values when constructing earnings measures. For example, if the day of the month that a job ended is missing, we will fill in a value of 15.
- For more substantive missing items, such as length of time in the job, we will impute based on job characteristics and other relevant follow-up and baseline data using hot-decking procedures.

Approach to secondary analysis

In addition to estimating impacts on secondary outcomes, the secondary analyses will also include (1) Bayesian analysis, (2) pooling data across programs and estimating impacts for all the programs together, and (3) examining the effects of the 2019 novel coronavirus disease (COVID-19) pandemic.

BAYESIAN ANALYSIS

To help readers interpret the findings, we will complement our main reporting of statistical significance (the frequentist analysis) with a Bayesian analysis. The Bayesian analysis will provide a probability that the true effect of the program is positive or greater than a specified amount—this is nuanced information that is helpful to practitioners and policymakers rather than just a conclusion that the program is probably effective or not. The Bayesian analysis also guards against the frequent misunderstanding about the meaning of statistical significance, which can lead to serious misinterpretation of study findings. Many people misinterpret statistical significance (p -value < 0.05) to mean that there is at most a 5 percent chance that the program had no effect rather than the correct conclusion that when the true effect is zero, there is a 5 percent chance that the impact estimate is statistically significant. A statistically significant impact does not necessarily imply a high probability that the program had an effect. Similarly, a lack of statistical significance does not necessarily mean that there is a low probability a program had an effect. The consequences of misinterpreting p -values can be so severe that several researchers have urged the field to abandon the use of p -values and statistical significance (Cooper et al. 2009; Gelman et al. 2013).

Overview of the BASIE approach

We plan to present findings from a Bayesian approach known as BASIE (BAyesian Interpretation of Estimates) (Deke and Finucane 2019).² We will apply this approach to estimate the *probability* that coaching had an effect of more than a specific amount on key study outcomes, rather than an indication of whether coaching had an effect at all. This approach applies Bayesian methods, drawing on both the effect directly estimated from the study's data and prior evidence about how common it is for programs to have effects.

The BASIE approach directly estimates the probability that the true effect of a program is a certain size. For example, we could draw conclusions about the likelihood that the impact is positive, such as, “There is a 75 percent chance that the program had a positive effect on average monthly earnings.” In addition, we could draw conclusions about the probability that the program had a large effect that readers are likely to regard as meaningful, such as, “There is a 50 percent chance that the program boosted average monthly earnings by \$250 or more.”

How the BASIE approach compares to other Bayesian methods. The BASIE approach differs from how researchers often apply Bayesian methods in two key ways. First, a common concern with Bayesian methods is that they can be subjective. Instead of drawing on prior evidence, they sometimes rely on prior beliefs about the effects of a program (Cooper et al. 2009). The BASIE framework avoids this concern by drawing only on prior evidence from similar evaluations, rather than on the researcher's beliefs about the programs' effects. Second, under the standard Bayesian approach, researchers often report only the Bayesian shrunken estimate (which is a weighted average of

² The components of BASIE draw on guidance from many sources (Gigerenzer and Hoffrage 1995; Gelman and Weakliem 2009; Gelman 2001, 2012, 2015a, 2015b, 2016; Gelman and Shalizi 2013).

the traditional effect estimate and prior evidence). In contrast, the BASIE approach encourages researchers to report both the main impact estimate (based only on study data) and the Bayesian shrunken estimate (Cooper et al. 2009; Gelman et al. 2013).

Information required to implement BASIE. The BASIE approach requires information that will come from our main analysis and additional information from other sources. In particular, the approach requires (1) the effect estimate and standard error, which we will estimate in our main analyses; and (2) prior information on how common it is for generally similar programs to have effects. The additional prior information will allow us to quantify how common it is to achieve effects of different sizes, such as how common it is to achieve positive effects or effects greater than a particular size.

Guidelines to selecting prior information

BASIE applies five guidelines for selecting and analyzing prior information: (1) use evidence from past evaluations, as opposed to beliefs about the effectiveness of programs that are not based in evidence; (2) select prior evidence that meets systematic standards for quality, such as studies reviewed by evidence clearinghouses; (3) statistically adjust evidence for variation in precision and possible biases that arise from how effects are reported; (4) select evidence that is similar to the programs and populations in the evaluation; and (5) examine and report sensitivity of findings to the selection of prior evidence.

There is not a general guideline for the number of past evaluations to assess. We will aim for 30 studies for this analysis. It is not necessary that the evidence we draw on for this work come from evaluations of coaching programs, so long as we are able to clearly articulate what the evidence represents. For example, the evidence could represent programs intended to help low-income people improve their employment outcomes.

Recommended source of prior information

We will base the priors on Pathways to Work Evidence Clearinghouse (Pathways), a project of the Office of Planning, Research and Evaluation (OPRE). Pathways aligns closely with the Employment Coaching evaluation because it focuses on studies of employment and training interventions for populations with low incomes. As described in more detail below, Pathways has enough studies to enable us to form prior information. In addition to Pathways, we also considered OPRE's Employment Strategies for Low-Income Adults Evidence Review (ESER) and the Department of Labor's Clearinghouse for Labor Evaluation and Research (CLEAR). We recommend focusing on Pathways rather than on ESER or CLEAR for two main reasons:

- 1. Pathways includes and builds on the research covered by ESER.** Therefore, we would not need to use ESER as a separate source of information.
- 2. Unlike Pathways, CLEAR does not focus exclusively on populations with low incomes or employment and training programs, so many of the studies from CLEAR would be less aligned with the Employment Coaching evaluation.** In addition, we anticipate that many of the studies from CLEAR that do focus on employment programs for populations with low incomes would also overlap with Pathways. Therefore, including CLEAR would offer little added benefit.

Selecting priors from Pathways

In advance of conducting the analyses, we plan to identify (1) a set of prior information that we will use for the Bayesian analysis to be presented in the body of the report and (2) other sets of prior information that we will use for sensitivity analyses to be presented in the appendix. We will identify the prior information using parameters in the Pathways database, including the quality of the study, the populations served by the intervention, and the intervention services. Table 8 details how we would use these parameters to define the priors for the main analysis and the sensitivity analyses. We recommend basing the priors for the main analysis on studies that are rated as high quality. However, we do not recommend limiting them further to take into account whether the intervention serves particular populations or provides particular services; because all studies in Pathways focus on employment and training programs for low-income populations, we view them as relevant prior evidence for this study. This choice also enables us to use a larger set of evidence as the basis for the priors and will be simple to describe.

Table 8.
Recommendations
for using
Pathways to
define prior
information

Parameter and description	Recommendation for main analysis	Recommendation for sensitivity analyses
Quality of study. Pathways rates the quality of evidence as either high, moderate, or low.	High-quality studies. We recommend focusing on studies that were rated as high quality. Because we expect that the Coaching Employment evaluation will meet the standards for a high-quality evaluation, focusing on high-quality studies will ensure a more comparable set of priors.	None. We considered conducting a sensitivity check using a sample that would additionally include outcomes that were rated as moderate quality. However, no studies in the Pathways database fall into this category, so a separate analysis would add little value.
Population. Pathways categorizes populations served by interventions based on eight dimensions: (1) whether they are cash assistance recipients, (2) whether they are disconnected or discouraged workers, (3) education level, (4) employment status and income level, (5) sex, (6) parental status, (7) whether they have specific employment barriers, and (8) whether they are young adults.	All populations. Because Pathways focuses on studies that include focal low-income populations—like the Employment Coaching programs in this study—we recommend including all focal populations in our main analysis.	Varies by program. For FaDSS and Goal4 It!, the two programs that have TANF participation as eligibility criteria, we recommend also conducting analyses focusing on cash assistance recipients. We do not recommend sensitivity analysis for the other programs.
Intervention services. Pathways categorizes intervention services into ten broad categories: (1) case management, (2) education, (3) employment retention services, (4) financial incentives, (5) health services, (6) sanctions, (7) supportive services, (8) training, (9) work and work-based learning, and (10) work readiness activities. A single intervention can include multiple services.	All service types. We recommend including interventions that offer all types of services in our main analysis. Given that Pathways focuses on employment programs, evaluations of all the intervention types can provide valid prior information.	Interventions that include one-on-one assistance. Because the coaching models can be developed to replace other one-on-one assistance models, such as case management, we recommend conducting a sensitivity analysis using only interventions that include one-on-one assistance.

Outcome domains and timing

We suggest using outcomes from prior studies that align with the Employment Coaching study’s confirmatory outcomes in terms of domains and timing of measurement:

- **Outcome domains.** Pathways categorizes outcomes into four possible categories: benefit receipt, education and training, earnings, and employment. We suggest focusing on earnings and benefit receipt because these are confirmatory outcomes in this evaluation. Pathways does not include information on self-regulation skills or economic hardship, the other confirmatory outcomes domain in this study. Thus, we plan to exclude self-regulation skills from the Bayesian analyses.
- **Timing of outcome measurement.** Pathways also categorizes the timing of measurement of outcomes as either “short-term” (18 months or less after participants are first offered services); “long-term” (between 18 months and 5 years after participants are first offered services); or “very long-term” (more than 5 years after participants are first offered services). We suggest focusing on categories that align with the timing of our data collection. For analyses of the first follow-up survey, we suggest focusing on short-term outcomes. For analyses of the second follow-up survey, we suggest focusing on long-term outcomes.

Expected sample sizes

Our proposed prior definitions will yield a sufficiently large sample of outcomes and studies for each outcome domain and time period we will include in the Bayesian analysis (Table 9). For each domain and time period, we expect to have 51 or more studies, exceeding our target of at least 30. We expect that our proposed sensitivity analyses will be adequately powered.

Table 9.
Estimated number of outcomes and studies available to form priors for the main analysis

Outcome domain	Time period	Anticipated sample size	
		Outcomes	Studies
Earnings	Short-term	69	51
	Long-term	72	59
Benefits receipt	Short-term	266	83
	Long-term	276	87

Source: Pathways database. The Bayesian analyses will be based on estimates of impacts in effect-size units. The anticipated number of studies represents the number of distinct studies that have at least one estimate in effect-size units for an outcome that falls in the outcome domain. The number of outcomes represents the number of available estimates in effect-size units for outcomes in the domain across all studies.

Presentation of Bayesian results

For the confirmatory outcomes, we plan to present the Bayesian results next to the main impact estimates (see Table 10 for an example). For each outcome domain, we will present several cutoffs that will allow us to draw conclusions about the probability that the true impact is greater than or less than the cutoffs. We will use the main estimates to draw conclusions about the effectiveness of the program and use the Bayesian findings to provide a complementary interpretation.

Table 10.
Example of
presentation of
Bayesian results

Outcome	Program group	Control group	Estimated impact	Probability that the true impact is:		
				Greater than 0	Greater than 100	Greater than 250
Average monthly earnings (survey; \$)						
Average monthly earnings (administrative; \$)						
Note: Mean values by research group and estimated impacts are derived from a regression model controlling for key baseline characteristics. Probabilities of true impact size are derived from Bayesian analysis that incorporate prior findings from the Pathways to Work Evidence Clearinghouse.						

POOLING ACROSS PROGRAMS

To examine the average effect of coaching programs in the Employment Coaching evaluation, we will estimate impacts pooled across all programs as part of the secondary analysis. Impact estimates from this analysis will be more precise than the program-specific estimates because of the larger sample size. In calculating pooled impact estimates, we will weight program-level impacts equally rather than in proportion to the size of their sample. Weighting each program according to the size of its sample would arbitrarily give some programs more importance when computing a pooled estimate. In contrast, weighting programs equally generates a more policy-relevant parameter: the impact observed for an average program in the evaluation, recognizing that each program represents a different implementation of employment coaching.

Each pooled estimate will be based on a regression model that controls for various baseline characteristics. All covariates will be interacted with binary variables identifying each program. This approach allows the influence of each explanatory variable to differ for each program and enables us to account for the fact that the baseline data sources are different for the MyGoals programs than for the other programs. The pooled models will use the same baseline covariates discussed in the previous section.

ANALYSIS OF THE COVID-19 PANDEMIC

The COVID-19 pandemic led to widespread lockdowns and other disruptions in the United States beginning in March 2020. The pandemic has had profound impacts on the operations of the employment coaching programs and on the broader economic context. All the programs continued operating, but they had to adapt the types of services provided and the means of providing them in many ways in response to social distancing requirements, decreased economic activity, and changing participant needs. All programs changed from mainly in-person to virtual interactions through at least spring 2020. We will conduct secondary and some exploratory analyses to assess how program impacts might have been influenced by the pandemic.

The pandemic did not disrupt the enrollment period for the programs in the Employment Coaching evaluation, but it could have affected service receipt for some program group participants. The pandemic began at least four months after the last study

enrollment for all programs (Table 11). Some program group participants would have still been receiving program services at that time, particularly those who enrolled toward the end of the enrollment period.

Table 11.
The timing of
the COVID-19
pandemic
relative to study
enrollment, by
program

Program	Last month of study enrollment	Number of months between last study enrollment and onset of the pandemic in March 2020
FaDSS	November 2019	4
Goal4 It!	November 2019	4
LIFT	November 2019	4
MyGoals Baltimore	September 2019	6
MyGoals Houston	July 2019	8

The pandemic occurred during the first follow-up period for some study participants in all programs, and during the second follow-up period for nearly all study participants. For example, the 9-month follow-up period would include the first pandemic-affected month (March 2020) for study participants at FaDSS, Goal4 It!, and LIFT who enrolled in June 2019 or later.

It is possible that the pandemic influenced program impacts. The direction of this influence is ambiguous. For example, if the programs were successful in improving goal-setting and self-regulation skills, those skills might have helped program group members better adapt, relative to control group members, to uncertainty in the post-pandemic economy, leading to larger earnings impacts than would be found under more typical economic conditions. Alternatively, the lack of jobs in the post-pandemic economy could have eliminated work opportunities that would have otherwise been available to program group members, leading to smaller program impacts on earnings than would be found under more typical economic conditions. Furthermore, lack of employment could lead to reduced opportunities to practice and maintain goal-setting and self-regulation skills in an employment setting, which could cause impacts on goal-setting and self-regulation skills to decay. Lack of employment also most likely affected receipt of public benefits, as did modifications to criteria for receiving benefits made in response to the pandemic. In addition, we do not know whether virtual provision of services is more or less effective than in-person service provision, or whether the shift to provide other types of services during the pandemic could influence impacts.

We will present some findings from this analysis in the main report. More detailed findings will be presented in an appendix or a separate report.

We propose two main analyses of the first follow-up data to assess how the pandemic might have affected impacts: (1) subgroup analysis based on date of the follow-up period and (2) impact analysis of earnings and public benefit receipt based on calendar month. We will determine the most appropriate pandemic-related exploratory analysis of the second follow-up data in accordance with how pandemic conditions evolve before the second follow-up data collection concludes in 2022.

Subgroup analysis based on date of follow-up period

We will conduct impact analysis on subgroups intended to isolate the impacts of the program both before and during the pandemic.

Impacts of the program before the pandemic. To examine impacts of the program before the pandemic, we will estimate impacts for the subgroup of study participants whose follow-up period ended before the pandemic began:

- For FaDSS, Goal4 It!, and LIFT, this sample will include study participants enrolled in March 2019 or earlier. The 9-month follow-up period for those enrolled in March 2019 runs through the end of 2019. Study participants who enrolled after that time would have a follow-up period for administrative earnings that includes the first quarter of 2020, when the pandemic began.
- For the MyGoals programs, this sample will include study participants enrolled in December 2018 or earlier. The 12-month follow-up period for those enrolled in December 2018 runs through the end of 2019.

The confirmatory goal-setting and attainment outcome and the confirmatory economic hardship outcome are measured at the time of the follow-up surveys. While the first follow-up period will have ended for all these participants before March, it may be that the participant did not respond to the survey until months after the end of the follow-up. For that reason, this analysis will not include survey responses that occurred during March 2020 or later. We do not need to exclude any cases from analysis of the confirmatory outcomes that use data on earnings from survey sources or data from administrative records (earnings from NDNH data and, for FaDSS and Goal4 It!, TANF benefit receipt). This is because this information covers the full follow-up period from the time of study enrollment rather than just the time of the survey response. We can obtain data over the same time period irrespective of when the survey was completed.

For all programs, sample members in these study enrollment cohorts represent at least half of all study enrollees (Table 12).

Table 12.
The timing of
the COVID-19
pandemic
relative to study
enrollment, by
program

Program	Study sample for which first follow-up ended before March 2020	Study sample for which last three months of first follow-up began after March 2020
FaDSS	482	183
Goal4 It!	354	123
LIFT	368	308
MyGoals (both locations)	1,079	289
Baltimore	475	196
Houston	604	93

Impacts of the program during the pandemic. To examine impacts of the program during the pandemic, we would like to estimate impacts on a sample whose confirmatory outcomes were measured entirely during the pandemic. However, there are no sample members whose full follow-up period occurred during the pandemic. The earnings outcomes in the confirmatory analysis use the full follow-up period as their reference period. Thus, this analysis will focus on earnings outcomes that use the last quarter of the follow-up period as the reference period so that we can examine impacts exclusively during the pandemic.

We will estimate impacts for study participants whose reference period for outcomes included in the analysis began after March 2020:

- For FaDSS, Goal4 It!, and LIFT, this sample will include study participants enrolled in September 2019 or later. The nine-month follow-up period for these cases ended in June 2020 or later. The last three months of the follow-up period began in April 2020 or later.
- For the MyGoals programs, this sample will include study participants enrolled in June 2019 or later. The 12-month follow-up period for these cases ended in June 2020 or later. The last three months of the follow-up began in April 2020 or later.

The total of sample members in all programs who meet these criteria is less than 300, a substantially smaller number than that included in analysis of program impacts before the pandemic. We will note this caveat when discussing these findings.

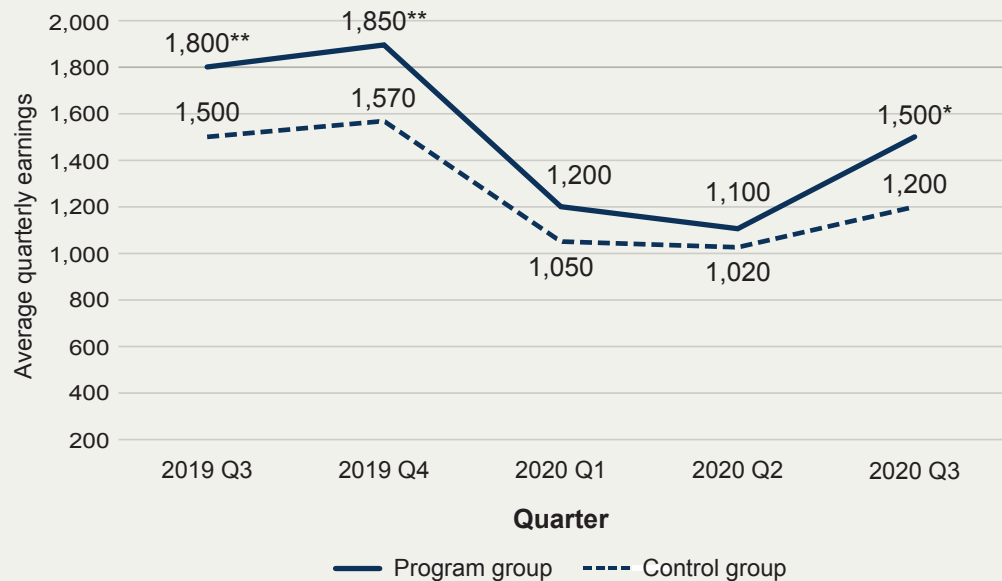
Impact analysis by calendar date

If the pandemic influenced program impacts, we would expect those effects to emerge in March 2020 and evolve over time in response to changing pandemic conditions. It is difficult to observe this type of effect with outcomes that are defined relative to the month of study enrollment, as are the main measures in this study and in others commonly used in program evaluation. Another option is to examine impacts on outcomes by calendar date. For example, using NDNH data, we can calculate the mean earnings in Quarter 1 of 2019 separately for all program and control group members. Comparing these means provides an impact estimate because any difference in the means can be attributed to the fact that one group was randomly assigned to receive program services and the other was not. Although this estimate will not be identical to our main impact estimate, it is likely to be similar if impacts are relatively constant over time.

Examining impacts by calendar date can show how the impacts changed in response to pandemic conditions. We will examine average earnings from NDNH data by calendar quarter from Quarter 3 of 2019 through Quarter 3 of 2020. We will also examine average TANF benefit receipt by month for this period. Figure 1 provides an illustrative example of what the earnings analysis might look like if (1) the program had impacts of about \$300 on average monthly earnings before the pandemic; (2) both program and control group earnings decreased sharply in response to the pandemic beginning in late Quarter 1 2020; and (3) average monthly earnings for the program group increased more rapidly than they did for the control group as pandemic conditions evolved.

We can conduct similar analyses using survey data. The survey job grid enables us to construct earnings estimates for any calendar month between the time of study enrollment and the time of survey response. We can pool all survey respondents whose job grid includes a given calendar month to get a mean earnings estimate for that month. We will construct such estimates separately for program and control group members for each month from August 2019 through October 2020.

Figure 1.
Illustrative
example of
average monthly
earnings from
administrative
records, by
calendar quarter
and research
group status



Source: The illustrative example is not based on actual data.

***/**/* Difference in earnings was statistically significant at the .01/.05/.10 levels, respectively, two-tailed test.

Approach to Exploratory Analysis

In addition to estimating impacts on exploratory outcomes, the exploratory analyses will also include (1) estimates of impacts by subgroup, (2) estimates of impacts on participants, (3) mediation analysis, and (4) robustness checks.

SUBGROUP ANALYSIS

As part of the exploratory analysis, we will check the consistency of impacts on confirmatory outcomes across subgroups. Examining differences in impacts by groups of participants, identified by their characteristics at study enrollment, can help our understanding of the magnitude of impacts as well as help programs think through whether there is a need to revise services.

While arguments can be made for why impacts might differ between many groups of participants owing to their characteristics at baseline, there is no hypothesis or expectation that the impacts would differ significantly by subgroup. Thus, all the subgroup analysis is part of the exploratory analysis. We will present findings from the subgroup analysis in the technical appendix, referencing them in the text of the main report as appropriate (per Table 1).

We will estimate subgroup analysis pooled across all programs when the relevant baseline information is available for all programs (and across FaDSS, Goal4 It!, and LIFT when the relevant information is not available for MyGoals). We will also estimate subgroup analysis separately for each program. Because the set of baseline information we have on study participants differs for MyGoals and the other programs, the subgroups we can examine differ across programs.

For adequate statistical power, we will not estimate separate impacts for subgroups with fewer than 300 study participants. A sample of 300 study participants would provide a minimum detectable effect size of 0.27 for outcomes based on administrative records and 0.32 for those based on survey reports. This requirement means that we can examine subgroups that represent 7 percent of the sample in analysis that pools across all programs, 17 percent for the MyGoals program (pooled across Houston and Baltimore locations), and about 35 percent for FaDSS, Goal4 It!, and LIFT. Thus we would not examine less common subgroups, especially in the program-specific analysis for FaDSS, Goal4 It!, and LIFT. For example, we will not include sex as a subgroup, as more than 80 percent of participants in each program consider themselves female. Similarly, we would not estimate the impacts for subgroups of participants who received TANF prior to study enrollment in FaDSS or Goal4 It!, as all program participants must receive TANF to be eligible for the coaching.

Table 13 presents some subgroups we will examine and the data source to be used to identify subgroups. We may expand the set of subgroups included in the exploratory analysis as needed to investigate patterns in the confirmatory analysis for each program.

Table 13.
Subgroups
to include in
exploratory
analysis

Subgroup	Program	Data source
Demographic and socioeconomic characteristics		
Age	All	Baseline survey (FaDSS, Goal4 It!, LIFT) MyGoals baseline form (MyGoals)
Race/ethnicity	FaDSS, Goal4 It!, LIFT Not in MyGoals because about 95 percent of study participants are Black, Non-Hispanic	Baseline survey
Number of children	All	Baseline survey (FaDSS, Goal4 It!, LIFT) Public Housing Authority administrative data (MyGoals)
Disability	MyGoals only Information is not available for other programs	Public Housing Authority
Received income from any public assistance or social insurance program	LIFT and MyGoals Not for FaDSS and Goal4 It!, as all FaDSS and Goal4 It! participants are TANF recipients	Baseline survey (LIFT) Public Housing Authority administrative data (MyGoals)

Table 13.
Subgroups
to include in
exploratory
analysis
(continued)

Subgroup	Program	Data source
Education, employment, and goal-setting		
Challenges to employment at baseline	FaDSS, Goal4 It!, LIFT Information is not available for MyGoals	Baseline survey (FaDSS, Goal4 It!, LIFT)
Education	All	Baseline survey (FaDSS, Goal4 It!, and LIFT) MyGoals baseline form (MyGoals)
Recent employment history at time of study enrollment	All	Baseline survey (FaDSS, Goal4 It!, and LIFT) Public Housing Authority administrative data (MyGoals) NDNH data for NDNH outcomes (All programs)
Goal-setting at baseline	FaDSS, Goal4 It!, and LIFT Information is not available for MyGoals	Baseline survey
Community		
Degree of urbanity	FaDSS only Insufficient variation in locations of other programs	County of FaDSS location
Whether in New York City, Los Angeles, or Chicago	LIFT only	Program data

ESTIMATES OF THE IMPACTS FOR THOSE WHO RECEIVED PROGRAM SERVICES

Our main impact analysis will compare outcomes for those assigned to the program group and those assigned to the control group, and provide estimates of the “intent to treat” (ITT) impact. However, policymakers and program administrators are also interested in estimates of the impact of the intervention on those who actually participated in the intervention—the “treatment on the treated” (TOT) impact. To estimate the TOT impact, we will apply the Bloom adjustment (Bloom 1984), which involves dividing the ITT estimate by the percentage of the program group who received intervention services. This approach is valid if no members of the control group receive coaching. We will estimate Bloom-adjusted impacts for the MyGoals programs, where about 7 percent of program group participants had no contact with the program in the nine months after study enrollment. We will not estimate TOT impacts for FaDSS, LIFT, or Goal4 It! because all or nearly all program group members received some coaching services. We might also conduct analyses to estimate impacts for those who received certain amounts or types of program services if doing so would be useful for interpreting findings from the confirmatory analysis.

MEDIATION ANALYSIS

As part of the exploratory analysis, we will conduct a mediation analysis, which will shed light on the mechanisms through which impacts emerge. We will use a two-step procedure to estimate this decomposition (Heckman and Pinto 2015; Kautz and

Zanoni 2015). This analysis will focus on outcomes with statistically significant impacts in the confirmatory analysis. For example, if there are statistically significant impacts on average monthly earnings, we could decompose the overall program impact on earnings into (1) a component attributable to impacts on potentially relevant outcomes like self-regulation, goal-setting, and hard skills; and (2) a component attributable to changes in other, unmeasured variables. In another example, if there is a statistically significant impact on goal-setting and attainment, we could decompose the overall program impact into (1) a component attributable to impacts on service-receipt outcomes such as receipt of one-on-one job assistance related to goal setting and attainment; and (2) a component attributable to changes in other, unmeasured variables.

We will use findings from the confirmatory analysis to determine the specific outcomes to include in the mediation analysis. We will not conduct this analysis if there are no statistically significant impacts in the confirmatory analysis. If conducted, this analysis will include intermediate outcomes for which there are statistically significant impacts.

ROBUSTNESS CHECKS

As part of the exploratory analysis, we will conduct analytic robustness checks to verify that the findings from our confirmatory analysis are not overly sensitive to specific analytic decisions that we made. We will perform robustness checks by re-running the confirmatory analyses using different specifications. For example, we will compare the findings of our confirmatory analyses from when we apply weights and when we do not, and we will compare the findings of our confirmatory analyses from when we use regression models with and without covariates. We will present findings from these sensitivity analyses in the technical appendix, referencing them in the text of the main report, as appropriate.

References

Barnow, B. S. and Greenberg, D. “Do Estimated Impacts on Earnings Depend on the Source of the Data Used to Measure Them? Evidence from Previous Social Experiments.” *Evaluation Review*, vol. 39, no. 2, 2015, pp. 179–228.

Benjamini, Y., and Y. Hochberg. “Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing.” *Journal of the Royal Statistical Society, Series B (Methodological)*, vol. 57, no. 1, 1995, pp. 289–300.

Bloom, H.S. “Accounting for No-Shows in Experimental Evaluation Designs.” *Evaluation Review*, vol. 8, no. 2, April 1984, pp. 225–246.
doi:10.1177/0193841X8400800205.

Cooper, H.M., L.V. Hedges, and J. Valentine (eds.). *The Handbook of Research Synthesis and Meta-Analysis*, 2nd edition. New York: Russell Sage Foundation, 2009.

Cox, D.R. *Analysis of Binary Data*. London: Chapman and Hall/CRC, 1970.

Deke, J., and M. Finucane. “Moving Beyond Statistical Significance: The BASIE (BAyeSian Interpretation of Estimates) Framework for Interpreting Findings from Impact Evaluations.” OPRE Report 2019-35. Washington, DC: U.S. Department of Health and Human Services, Administration for Children and Families, Office of Planning, Research, and Evaluation, 2019.

Deke, J., and M. Puma. “Coping with Missing Data in Randomized Controlled Trials.” Evaluation Technical Assistance Brief, no. 3. Washington, DC: Administration for Children and Families, Office of Adolescent Health, 2013.

Gelman, A. “Induction and Deduction in Bayesian Data Analysis.” *Rationality, Markets and Morals*, vol. 2, 2001, pp. 67–78.

Gelman, A. “Ethics and Statistics: Ethics and the Statistical Use of Prior Information.” *CHANCE*, vol. 25, no. 4, 2012, pp. 52–54.

Gelman, A. “Prior information, not prior belief.” Statistical Modeling, Causal Inference, and Social Science blog, 2015a. Available at <http://andrewgelman.com/2015/07/15/prior-information-not-prior-belief/>. Accessed June 7, 2018.

Gelman, A. “The General Problem I Have with Noninformatively-Derived Bayesian Probabilities Is That They Tend to Be Too Strong.” Statistical Modeling, Causal Inference, and Social Science blog, 2015b. Available at <http://andrewgelman.com/2015/05/01/general-problem-noninformatively-derived-bayesian-probabilities-tend-strong/>. Accessed July 26, 2021.

Gelman, A. “What Is the ‘True Prior Distribution’? A Hard-Nosed Answer.” Statistical Modeling, Causal Inference, and Social Science blog, 2016. Available at <http://andrewgelman.com/2016/04/23/what-is-the-true-prior-distribution-a-hard-nosed-answer/>. Accessed July 26, 2021.

- Gelman, A., and C.R. Shalizi. "Philosophy and the Practice of Bayesian Statistics." *British Journal of Mathematical and Statistical Psychology*, vol. 66, no. 1, 2013, pp. 8–38. doi:10.1111/j.2044-8317.2011.02037x
- Gelman, A., and D. Weakliem. "Of Beauty, Sex and Power." *American Scientist*, vol. 97, no. 4, 2009, pp. 310–316.
- Gelman, A., J.B. Carlin, G.S. Stern, D.B. Dunson, A. Vehtari, and D.B. Rubin. *Bayesian Data Analysis*, 3rd edition. Boca Raton, FL: CRC Press, 2013.
- Gigerenzer, G., and U. Hoffrage. "How to Improve Bayesian Reasoning Without Instruction: Frequency Formats." *Psychological Review*, vol. 102, no. 4, 1995, pp. 684–704.
- Heckman, J.J., and R. Pinto. "Econometric mediation analyses: Identifying the sources of treatment effects from experimentally estimated production technologies with unmeasured and mismeasured inputs." *Econometric Reviews*, vol. 34, no. 1-2, 2015, pp. 6–31.
- Hedges, L.V. "Distribution Theory for Glass's Estimator of Effect Size and Related Estimators." *Journal of Educational Statistics*, vol. 6, no. 2, Summer 1981, pp. 107–128.
- Kautz, T., and W. Zanoni. "Measuring and Fostering Non-Cognitive Skills in Adolescents: Evidence from Chicago Public Schools and the OneGoal Program." University of Chicago, Department of Economics, 2015.
- Moore, Q., S. McConnell, A. Werner, T. Kautz, K. Joyce, K. Borradaile, and B. Boland. "Evaluation of Employment Coaching for TANF and Related Populations: Evaluation Design Report." Report submitted to the U.S. Department of Health and Human Services, Administration for Children and Families. Washington, DC: Mathematica, August 2019.
- Orr, Larry L. *Social Experiments: Evaluating Public Programs with Experimental Methods*. Thousand Oaks, CA: Sage, 1999.
- Puma, M.J., R.B. Olsen, S.H. Bell, and C. Price. "What to Do When Data Are Missing in Group Randomized Controlled Trials." NCEE 2009-0049. Washington, DC: National Center for Education Evaluation and Regional Assistance, October 2009.
- Roth, R.M., P.K. Isquith, and G.A. Gioia. "Behavior Rating Inventory of Executive Function—Adult Version (BRIEF-A)." *Archives of Clinical Neuropsychology*, vol. 20, no. 7, October 2005.
- Schochet, P.Z. "An Approach for Addressing the Multiple Testing Problem in Social Policy Impact Evaluations." *Evaluation Review*, vol. 33, no. 6, December 2009, pp. 539–567. doi:10.1177/0193841X09350590.

Appendix A: Psychometric analysis of outcomes in the goal-setting and self-regulation skill domain

We conducted an analysis to determine how to define measures of goal-setting and self-regulation skills using the responses from questions on the surveys, and to assess whether these measures are reliable and valid. We had previously identified four measures through a psychometric analysis of the data collected on the baseline survey: (1) goal-setting; (2) self-esteem (Rosenberg 1965); (3) emotional control and self-monitoring (Roth et al. 2005); and (4) task monitoring, planning, and initiation (Roth et al. 2005).

This appendix describes a similar analysis we conducted using responses to items in the first follow-up survey. Our analyses proceeded in three steps. First, we posited how to group individual survey items into measures of specific self-regulation skills. In this step, we developed new groupings of items on goals and self-regulation skills that were included in the follow-up surveys but not on the baseline survey. For measures that also appeared on the baseline survey, we used the groupings suggested by our analyses of the baseline survey. Second, we examined the reliability and validity of the measures using data from the first follow-up survey, including those that we previously had examined using data from the baseline survey. Third, informed by the results, we revised the measures slightly and confirmed the reliability and validity of the final, proposed version of the measures.

STEP 1. GROUPING ITEMS INTO MEASURES

In this step, we focused on grouping items that were on the first follow-up survey but not on the baseline survey. Three measures of self-regulation skills were developed from items common across the baseline and follow-up surveys. Because these three measures were based on existing instruments and performed well using data from the baseline survey, we did not revisit their definitions but did confirm their reliability and validity using data from the follow-up survey. The follow-up survey also included two sets of study-developed items that did not appear in the baseline survey. The new items were designed to capture aspects of goal-related skills and self-regulation skills as demonstrated in the context of employment.

Goal-related skills. In addition to the three goal-related items in the baseline survey, the follow-up survey included five other items related to goals (Table A.1). We considered two ways to group the items into scales: (1) an overall measure of goal-setting and attainment, and (2) separate measures of goal-setting and goal attainment (see the last column of Table A.1). Because we did not have prior evidence on these items, we tested both options to see which grouping fit the data better. As described in the next section, our analyses supported a single measure: goal-setting and attainment.

Self-regulation skills in the employment context. In addition to the one item in the baseline survey designed to capture self-regulation skills (as demonstrated in the

context of employment), the follow-up survey included five other items designed to capture those skills (Table A.2). We posited that these six items together would capture a single skill: employment-related self-regulation.

Table A.1.
Goal-related
items in the first
follow-up survey

#	Item	In baseline survey?	Goal-setting or goal attainment?
1	I know I need to get a job or a better job and really think I should work on finding one.	Yes	Goal-setting
2	I set employment goals based on what is important to me or my family.	No	Goal-setting
3	I set long-term employment goals that I hope to achieve (such as finding a job, finding a better job, getting promoted, or enrolling in further education).	Yes	Goal-setting
4	I set specific short-term goals that will allow me to achieve my long-term employment goals.	Yes	Goal-setting
5	Based on everything I know about myself, I believe I can achieve my employment goals.	No	Goal attainment
6	When I set employment goals, I think about barriers that might get in my way and make specific plans for overcoming those barriers.	No	Goal attainment
7	Even when I face challenges, I continue to pursue my employment goals.	No	Goal attainment
8	I keep track of my overall progress toward my long-term employment goals and adjust my plans if needed.	No	Goal attainment

Table A.2.
Items capturing
self-regulation
skills in the
employment
context in the first
follow-up survey

#	Item	In baseline survey?
1	Lost your temper with someone other than friends or family.	No
2	Said something that you later regretted to someone other than friends or family.	No
3	Decided not to apply for a job because you didn't think you would get an interview.	No
4	Overcame a barrier that could have prevented you from finding or keeping a job.	No
5	Been late for a job, interview, program meeting, class, or training session.	Yes
6	Missed an appointment related to work, looking for a job, a program, school, or training for a reason other than you were sick or ill.	No

STEP 2. EXAMINING THE RELIABILITY AND VALIDITY OF THE MEASURES

To examine whether the candidate measures performed well, we conducted two analyses. First, we estimated Cronbach's alpha for each of the five measures. This is a statistic that provides evidence on internal consistency reliability, the degree to which different items for a given measure produce similar results. Second, we conducted a confirmatory factor analysis that sheds light on aspects of validity—whether the measures capture what they were designed to measure. The confirmatory factor analysis examined (1) factor loadings

that capture the extent to which each item relates to the underlying self-regulation skill (related to internal consistency reliability); (2) model fit statistics that summarize whether the groupings of items into measures fit the data well overall (exhibit model validity); and (3) correlations between the pairs of skills that suggest whether separate measures capture different constructs (exhibit discriminant validity). We conducted the analysis twice, once with a single goal-setting and attainment measure, and once with separate measures of goal-setting and goal attainment. In both cases, the models also included the four other measures. We restricted the analysis sample to the control group to limit the possible perception that we selected definitions of outcome measures based on the results of the impact analysis. We assessed whether the hypothesized measures met standard criteria for reliability and validity by examining how well corresponding statistics compared to target values (Table A.3). As discussed in Kautz and Moore (2020), we viewed these criteria as guidelines, not strict rules.

Table A.3.
Criteria used
for assessing
reliability and
validity

Type of reliability or validity	Statistic	Target value
Internal consistency reliability	Cronbach's alpha	At least 0.65 (DeVellis 2017).
Internal consistency reliability	Factor loading	0.40 or above (Stevens 2012), particularly in cases when Cronbach's alpha is low and when the sign matches the theoretical relationship between the item and factor.
Model validity (overall model fit)	Root mean square error of approximation (Steiger and Lind 1980)	0.05 or below for a close fit and 0.08 or below for a reasonable fit, as Browne and Cudeck (1992) suggested on the basis of practical experience.
Model validity (overall model fit)	Comparative Fit Index (Bentler 1990a)	0.90 or above as suggested by Brown (2015) based on analysis by Bentler (1990b).
Model validity (overall model fit)	Tucker–Lewis Index (Tucker and Lewis 1973)	0.90 or above, as suggested by Brown (2015) on the basis of analysis by Bentler (1990b).
Discriminant validity	Correlation between factors	Less than 0.80 and they are theoretically distinct (Brown 2015).

All five of the measures met the criteria in Table A.3 with three exceptions:

- 1. A correlation between factors that exceeded the target value of 0.80.** We estimated that the correlation between the two separate measures of goal-setting and goal-attainment was 0.91. This estimate exceeds our cutoff for discriminant validity, suggesting that the two measures capture the same underlying skill. As an additional check, we used the Kaiser criterion (Kaiser 1960) to estimate the number of skills captured by the group of goal-related items and found that the items in the combined goal-setting and attainment measure captured a single skill. **Recommendation:** On the basis of this evidence, we propose using the combined goal-setting and attainment measure.
- 2. A factor loading that had an unexpected sign.** A factor loading ranges from -1 to 1 , and a positive factor loading indicates that higher values on the item are positively associated with the overall measure, whereas a negative factor loading

indicates that higher values on the item are negatively associated with the overall measure. The sign of the factor loading (negative or positive) should match the expected relationship with the skill. However, the factor loading had an unexpected sign for one item in the employment self-regulation measure (“Overcome a barrier that could have prevented you from finding or keeping a job”). The factor loading suggested that participants who overcame barriers more frequently had lower self-regulation skills. One possibility is that people who have higher levels of self-regulation skills experience fewer barriers to begin with, so they also report overcoming fewer barriers. **Recommendation:** On the basis of this evidence, we propose removing the item from the scale of employment self-regulation.

3. A factor loading that was lower than the target value of 0.40. The factor loading for one item in the goal-setting and attainment scale (“I know I need to get a job or a better job and really think I should work on finding one”) was 0.21. The downside of retaining an item with a low factor loading is that it could reduce the scale’s overall internal consistency reliability. However, when including this item, the Cronbach’s alpha of the goal-setting and attainment scale is 0.86, well above our target of 0.65. **Recommendation:** Because this item does not threaten the scale’s overall internal consistency reliability and was included in the baseline goal-setting scale, we recommend retaining it in the follow-up scale for consistency.

STEP 3. THE RELIABILITY AND VALIDITY OF THE REVISED MEASURES

After making our two proposed changes, we re-estimated the Cronbach’s alpha and the confirmatory factor model for the combined goal-setting and attainment measure, the revised employment self-regulation measure, and the three measures of self-regulation skills we had developed using items from the baseline survey. For all scales, Cronbach’s alpha met our criteria for internal consistency reliability (Table A.4). Aside from the factor loading for the one item in the goal-setting and attainment scale discussed in Step 2, the factor loadings exceeded the target factor loading of 0.40, with values ranging from 0.52 to 0.88. Similarly, the overall fit statistics and correlations between skills met our criteria for model validity and discriminant validity (Tables A.5 and A.6). Judging from the results of these analyses, we do not propose additional revisions to the measures. For the full list of items in each of the five measures, see Tables 2 and 4 in the main body of this technical supplement.

Table A.4.
Reliability of
measures of
self-regulation
skills

Measure	Cronbach's alpha	Meets criterion
Goal-setting and attainment ^a	0.86	Yes
Self-esteem ^b	0.66	Yes
Emotional control and self-monitoring ^c	0.88	Yes
Task monitoring, planning, and initiation ^c	0.92	Yes
Employment self-regulation ^d	0.66	Yes

Source: Evaluation of Employment Coaching first follow-up survey.

^a 0–3-point scale based on the extent to which respondents agree with statements that reflect a high level of goal-setting and attainment skills. The scale indicates whether they (0) strongly disagree, (1) disagree, (2) agree, or (3) strongly agree.

^b A 0–3-point scale based on the extent to which respondents agree with statements that reflect a high level of self-esteem. The scale indicates whether they (0) strongly disagree, (1) disagree, (2) agree, or (3) strongly agree.

^c A 0–2-point scale that indicates whether respondents have problems related to the skill (0) often, (1) sometimes, or (2) never.

^d A 0–3-point scale based on the frequency with which respondents exhibit behaviors that reflect a lack of employment self-regulation skills. The scale indicates whether they exhibit specific behaviors (0) a few times a week, (1) a few times a month, (2) about once a month, or (3) hardly ever or never.

Table A.5.
Model fit statistics
of measures of
self-regulation
skills

Model fit statistic	Value	Meets criterion
Root mean square error of approximation		
Estimate	0.047	Yes
95 percent confidence interval (lower and upper bounds)	0.045 to 0.049	Yes
Comparative Fit Index	0.963	Yes
Tucker–Lewis Index	0.960	Yes

Source: Evaluation of Employment Coaching first follow-up survey.

Note: The estimates for each sample come from a single confirmatory factor model that assumes five factors that correspond to the five self-regulation skills. The items corresponding to each skill are constrained to relate only to that skill. The factors are not constrained to be independent.

**Table A.6.
Correlations
between measures
of self-regulation
skills**

Skill 1	Skill 2	Correlation	Meets criterion
Goal-setting and attainment ^a	Self-esteem ^b	0.43	Yes
Goal-setting and attainment ^a	Emotional control and self-monitoring ^c	0.21	Yes
Goal-setting and attainment ^a	Task monitoring, planning, and initiation ^c	0.29	Yes
Goal-setting and attainment ^a	Employment self-regulation ^d	0.14	Yes
Self-esteem ^b	Emotional control and self-monitoring ^c	0.57	Yes
Self-esteem ^b	Task monitoring, planning, and initiation ^c	0.58	Yes
Self-esteem ^b	Employment self-regulation ^d	0.49	Yes
Emotional control and self-monitoring ^c	Task monitoring, planning, and initiation ^c	0.74	Yes
Emotional control and self-monitoring ^c	Employment self-regulation ^d	0.70	Yes
Task monitoring, planning, and initiation ^c	Employment self-regulation ^d	0.54	Yes

Source: Evaluation of Employment Coaching first follow-up survey.

Note: The estimates come from a single confirmatory factor model that assumes five factors that correspond to the five self-regulation skills. The items corresponding to each skill are constrained to load only on that skill. The factors are not constrained to be independent.

^a 0–3-point scale based on the extent to which respondents agree with statements that reflect a high level of goal-setting and attainment skills. The scale indicates whether they (0) strongly disagree, (1) disagree, (2) agree, or (3) strongly agree.

^b A 0–3-point scale based on the extent to which respondents agree with statements that reflect a high level of self-esteem. The scale indicates whether they (0) strongly disagree, (1) disagree, (2) agree, or (3) strongly agree.

^c A 0–2-point scale that indicates whether respondents have problems related to the skill (0) often, (1) sometimes, or (2) never.

^d A 0–3-point scale based on the frequency with which respondents exhibit behaviors that reflect a lack of employment self-regulation skills. The scale indicates whether they exhibit specific behaviors (0) a few times a week, (1) a few times a month, (2) about once a month, or (3) hardly ever or never.

References for Appendix A

- Bentler, P.M. "Fit Indexes, Lagrange Multipliers, Constraint Changes and Incomplete Data in Structural Models." *Multivariate Behavioral Research*, vol. 25, no. 2, 1990a, pp. 163–172.
- Bentler, P.M. "Comparative Fit Indexes in Structural Models." *Psychological Bulletin*, vol. 107, no. 2, 1990b, pp. 238–246.
- Brown, T.A. *Confirmatory Factor Analysis for Applied Research*, 2nd edition. New York: Guilford Publications, 2015.
- Browne, M.W., and R. Cudeck. "Alternative Ways of Assessing Model Fit." *Sociological Methods Research*, vol. 21, no. 2, 1992, pp. 230–258.
- DeVellis, R.F. *Scale Development: Theory and Applications*, 4th edition. Los Angeles: Sage Publications, 2017.
- Kaiser, H.F. "The Application of Electronic Computers to Factor Analysis." *Educational and Psychological Measurement*, vol. 20, no. 1, 1960, pp. 141–151.
- Kautz, T., and Q. Moore. "Selecting and Testing Measures of Self-Regulation Skills Among Low-Income Populations." OPRE Report #2020-138. Washington, DC: Office of Planning, Research, and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services, 2020.
- Rosenberg, M. *Society and the Adolescent Self-Image*. Princeton, NJ: Princeton University Press, 1965.
- Roth, R.M., P.K. Isquith, and G.A. Gioia. "Behavior Rating Inventory of Executive Function—Adult Version: Professional Manual." Lutz, FL: Psychological Assessment Resources, 2005.
- Steiger, J.H., and J.C. Lind. "Statistically Based Tests for the Number of Common Factors." Paper presented at the annual meeting of the Psychometric Society, Iowa City, IA, 1980.
- Stevens, J.P. *Applied Multivariate Statistics for the Social Sciences*, 5th edition. New York: Routledge, 2012.
- Tucker, L.R., and C. Lewis. "A Reliability Coefficient for Maximum Likelihood Factor Analysis." *Psychometrika*, vol. 38, no. 1, 1973, pp. 1–10.

